ACTIVE LEARNING FOR THE PREDICTION OF ASPARAGINE AND ASPARTATE HYDROXYLATION SITES ON HUMAN PROTEINS

By

Festus Omonigho Iyuke, B.Sc., M.Sc.

ACTIVE LEARNING FOR THE PREDICTION OF ASPARAGINE AND ASPARTATE HYDROXYLATION SITES ON HUMAN PROTEINS

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Masters of Applied Science

in Biomedical Engineering

Ottawa-Carleton Institute for Biomedical Engineering

Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario, Canada
September 2011

Your file Votre référence
ISBN: 978-0-494-83033-8
Our file Notre référence
ISBN: 978-0-494-83033-8

The undersigned recommend to the
Faculty of Graduate and Postdoctoral Affairs
acceptance of the thesis

ACTIVE LEARNING FOR THE PREDICTION OF ASPARAGINE AND ASPARTATE HYDROXYLATION SITES ON HUMAN PROTEINS

Submitted by Festus Omonigho Iyuke
in partial fulfillment of the requirements for
the degree of Master of Applied Science in Biomedical Engineering

Thesis Co-Supervisor	Thesis Co-Supervisor
Dr. James R Green	Dr. William G. Willmore

Dr. Howard Schwartz
Chair, Department of Systems and Computer Engineering

Abstract

This thesis reports on the development and evaluation of a pool-based active learning approach to create support vector machine (SVM) classifiers for the prediction of asparagine/aspartate (N/D) hydroxylation sites on human proteins. The verification of hydroxylation sites on human proteins in wetlab experiments is very costly and sometimes time-consuming to achieve. The active learning procedure could therefore be used to choose which putative hydroxylation sites should be selected for future wetlab experimental validation and verification in order to gain maximal information. Using a dataset of N/D sites with known hydroxylation status, we here demonstrate through simulations that active learning query strategies can achieve higher classification performance with fewer labelled training instances for hydroxylation site prediction, compared to traditional passive learning. The active learning query strategies (uncertainty, density-uncertainty, certainty) are shown to identify the most informative unlabelled instances for annotation by an Oracle at each learning cycle. Furthermore, our experimental results also show that active learning strategies are highly robust in the presence of class imbalance in the available training data.

Considering that simulations clearly demonstrated the advantage of active learning for this application, certainty-based and uncertainty-based strategies were therefore applied to select the most informative 20 putative N/D hydroxylation sites from the 1.3 million putative N/D hydroxylation sites in the entire human proteome. Only two of these

proteins were successfully isolated, quantified, and overexpressed in mammalian cells in an *in vitro* experiment, due to experimental limitations. The biological activity of these proteins was verified using Western blotting, immunoprecipitation, and Coomassie stain analysis based on the protein expression identified on an SDS-PAGE gel. The successful identification of these proteins' overexpression on the gel lays the foundations for the determination of the true annotation of these putative N/D hydroxylation sites via mass spectrometry. Following the active learning algorithm, ultimately, the classification of these new N/D sites will be used to further increase the prediction accuracy of the SVM-based classification model.

Acknowledgements

First and foremost, I would like to give a big thanks to my supervisors, Dr. James Green from the Department of Systems and Computer Engineering and Dr. William Willmore from the Biology Department for all your support during my thesis research. I do strongly appreciate your immeasurable scientific guidance and advice, especially your patience, suggestions and constructive supervision. Earnestly, I would like to thank you for all the invaluable input and unwavering support and for allowing me to learn some tools. More importantly, your enthusiasm in research has left me with deep appreciation for bioinformatics and biological sciences. A special thanks to Mrs. Zhen Liu, Drs. Green and Willmore's past student, who made the datasets available for the current study.

A special thanks to Dr. Green for steering me in the right direction with the schematic representation of active learning procedure and providing all required resources, tools and dataset used in the implementation of active learning algorithm. For training me in the background knowledge of machine learning and clear focus on active learning as relates to hydroxylation sites.

To Dr. Willmore, a special thanks to you too, you have provided me with the background in biochemistry and trained me in biochemistry laboratory techniques, other bioinformatics tools and provided many resources towards achieving the goals of this research. For allowing Dr. Green to bring me into this project which benefitted me greatly. For providing me a mentor during the wetlab experimental validation session, who not only mentored me but actually immersed himself with the actual wetlab experimental validation process, when I was going through the learning curve of mastering biochemistry lab techniques. Even after mastering some of the biochemistry

techniques, he was always there to answer any questions I might have. I would also like to thank Andrew Seal who provided much of the lab supervision on experimental protocol. The next big thanks goes to all other members of the Willmore Lab, including Arran McBride, Andrew Robinette, Kendra Young, Skye McBride and Trisha Mackie, who helped me a lot to become familiar with the wetlab tools and techniques. Without your assistance, the experimental validation component in Dr. Willmore's Lab would not have made it as far as it did.

I would earnestly like to thank Dr. Chih-Jen Lin and Chia-Hua Ho of LibSVM for providing me with sample code to tune LibSVM, and Dr. Settles Burr and other members of the Active Learning (Machine Learning) group on Google Groups. Thank you all for providing me with some useful insights and academic papers for the active learning query strategy.

I would to thank Dr. Diana Inkpen from the University of Ottawa, Dr. Richard Dansereau, and Dr. Andrew Marble for being my committee members, and giving me helpful suggestions and comments. I would also like to thank all my friends, lab mates who have shown a great deal of friendliness and support. I would like to thank the Natural Science and Engineering Research of Canada and the Department of Systems and Computer Engineering, Carleton University, for funding this research and my graduate studies.

Finally, I would like to thank all members of Iyuke's family for their support, love and encouragement. A very special thanks to Dr. Sunny Iyuke for your continuous and invaluable advice has been a great source of inspiration. To you, Bernadette Starrs, a warm and lovely thanks for your emotional support, encouragement, patience, understanding and steadfast love. It is highly appreciated. Last, but not least, to you

Success Iyuke, as well as your twin brother and sister; thanks for your understanding. Thank you all.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AA	Amino Acid
A_{260}	Absorbance at 260 nm
A_{320}	Absorbance at 320 nm
AAindex	Amino acid index
AHC	Agglomerative hierarchical clustering
APC	Affinity propagation clustering
AP2M1	AP2 complex, subunit MU, isoform B
ARD	Ankyrin repeat domain
Asn/N	Asparagine
AUC	Area under ROC curve
BSA	Bovine serum albumin
C	Classifier
C	Slack penalty coefficient
CAD	Carboxyl-terminal transcription activation domain
cAMP	Cyclic 3', 5'-adenosine monophosphate
CBAL	Class balanced active learning
CCBE1	Collagen- and calcium-binding EGF domain containing protein 1precursor
Conf_values	Confidence values
CREB	cAMP response element-binding
CTLA-4	Cytotoxin T-lymphocyte Antigen 4
D	Aspartate
DNA	Deoxyribonucleic acid
ExPASy	Expert Protein Analysis System
FIH	Factor inhibiting hypoxia-inducible factor
FN	False negative
FP	False positive
H	Hilbert space

HEK293	Human embryonic kidney 293 cells
HIF	Hypoxia-inducible factor
HRP	Horseradish perioxide
IDV	Integrated density value
ISC	Intrinsic stopping criteria
K	Lysine
kDa	Kilodalton
KNN	K-nearest neighbour
L	Labelled instances
LTBP	Latent transforming growth factor-β binding protein
M	Molar
MCC	Matthew correlation coefficient
mRNA	Messenger Ribonucleic acid
MS	Mass spectrometry
NCBI	National centre for biotechnology information
Neg	Negative class
NO/YES	Non-hydroxylated/hydroxylated sites
Num_Neg	Number of Negative
Num_Pos	Number of Positive
OH	Hydroxyl group
P	Proline
PAGE	Polyacrylamide gel electrophoresis
Pos	Positive class
PPP1R13L	RelA-associated inhibitor
Pred_label	Predicted labels
PPV	Predictive positive value
PSSM	Position-specific scoring matrix
PTM	Post translation modification
PVDF	Polyvinylidene difluoride
Q	Querying function
QBC	Query-by-committee

R	Arginine
RBF	Radial basis function
RKHS	Reproducing kernel Hilbert spaces
RNA	Ribonucleic acid
ROC	Receiver operating characteristics curve
SA	Surface accessibility
SDS	Sodium dodecylsulfate
SS	Secondary structure
SVM	Support vector machine
T	Thymine
TN	True negative
TP	True positives
TP53BP2	Apoptosis-stimulating protein of p53 isoform 1 or ASPP2
tRNA	Transfer Ribonucleic acid
U	Unlabelled instances
U	Uracil
V	Volts

1 INTRODUCTION

1.1 Introduction

Many proteins undergo some form of post-translational modification (PTM) or degradation after translation. Both reversible and non-reversible modifications are essential to a variety of biological processes in humans including signal transduction and enzyme activation/inactivation and may also lead to pathological changes and undesirable diseases (Lee, et al., 2006; Basu and Plewczynski 2010). Hydroxylation is an important example of oxygen-dependent protein post-translational modification. During the physiochemical process of hydroxylation, a protein amino acid side chain is modified by the attachment of at least one hydroxyl group (OH) (Hu et al., 2010) in a reaction catalyzed by enzymes known as hydroxylases. It is a reaction that is also dependent upon iron, ascorbic acid (vitamin C) and α-ketoglutarate. Hydroxylation is important for oxygen sensing in cells, and lack of hydroxylation signals low oxygen (hypoxic) conditions within the cells. A number of amino acids may be hydroxylated, including proline, lysine, tyrosine, tryptophan, phenylalanine, asparagine and aspartate, but this thesis will focus on the hydroxylation of asparagine/aspartate (N/D). The best studied protein which undergoes this type of PTM is the hypoxia-inducible factor (HIF) protein, a transcription factor that acts as major regulator in the detection and response to low oxygen in the human tissues (Peet & Linke 2006). Asparagine hydroxylation occurs adjacent to the carboxyl-terminal transcriptional activation domain (CAD) of HIF, repressing transactivation activity of HIF by blocking the interaction of the HIF CAD with the transcriptional coactivators protein p300 (i.e., the required protein for p53 gene expression) and preventing the activation of transcription. The asparagine hydroxylase in this case is Factor Inhibiting Hypoxia-inducible factor (FIH), an enzyme that has also been

known to catalyze the hydroxylation of highly conserved asparagine residues within the ubiquitous ankyrin repeat domain (ARD) of other proteins. For example, the hydroxylation of aspartate residues also occurs in the ARD of ankyrinR and is catalyzed by FIH (Peet et al., 2004; Cockman et al., 2009). Therefore, an accurate prediction model will help gain useful insights and understanding of the complex physiochemical mechanism of asparagine/aspartate hydroxylation sites.

Experimental identification of N/D hydroxylation sites is commonly performed by mass spectrometry (Cockman et al., 2009) which is resource-intensive, expensive, and time-consuming. This method can also suffer from high rates of false positive results; hence, supervised machine learning has been adopted to design a classifier to predict the hydroxylation sites, trained from a small set of previously annotated protein samples. Currently, there is only one supervised machine learning model for the prediction of N/D hydroxylation sites (Liu, 2009; a former student of Drs. Willmore and Green), who trained support vector machine (SVM) classifiers with a leave-one-out cross-validation test method and achieved a recall of 92.73% and the precision rate of 61.45% over a dataset of 55 confirmed positive N/D hydroxylation sites, and 1758 negative sites. Another supervised prediction model for the hydroxylation sites of proline and lysine utilized a nearest neighbour approach and was evaluated by jackknife cross-validation (Hu et al., 2010). Performance in terms of sensitivity, specificity and Matthew's correlation coefficient were 64.8%, 81.6%, and 0.461 respectively when evaluated on hydroxyproline dataset, and 70.4%, 88.0%, 0.592 respectively with a hydroxylysine dataset. Such PTM prediction models must often be trained on hundreds (even thousands) of annotated protein samples (instances) to achieve useful performance levels. However, gathering a dataset where all

instances are labelled as positive or negative can be very costly. Active learning addresses this intrinsic bottleneck, by allowing the active learner to iteratively and intelligently select which data points should be labelled for building the training dataset to learn the predictive model.

Active learning is an iterative approach to train the best classifier possible with the currently available labelled training data, and then using that classifier to select which instances should be labelled and added to the dataset to ultimately create a more accurate classifier. This careful way of choosing data points to be added to the training set will enable the learner to reach high performance using as few labelled data points as possible.

Procedure: Pool-Based Active Learning Process

Input: Randomly pick an initial small training set L, and a pool of unlabelled data set U

Use L to train the initial classifier C.

Repeat

- Use sampling strategies to select "*most informative*" unlabelled instances from the unlabelled pool U, and query Oracle for labelling
- Add newly labelled instance to L, and remove it from U
- Use L to retrain the current classifier C and evaluate the classifier's performance on an independent test set
- Use current classifier C to label all unlabelled points in U

Until the predefined stopping criteria is reached or all unlabelled instances have been selected

---

Figure 1.1. Generic Pool-Based Active Learning with Query Strategy

Pool-based active learning is illustrated above in Figure 1.1. and below in Figure 1.2. In an active learning cycle, a classifier C is trained on an initially small training set of labelled instances *L*. Then *n* new instances are chosen from pool of unlabelled instances *U,* according to the current classifier and querying strategy *Q* criterion. Thus, these queried instances are given to the Oracle for labelling. The Oracle is able to label or annotate any instance with its correct class label for a fixed cost. In our case, the Oracle will ultimately be a wetlab experimentalist who can determine whether an N/D site on a protein is a hydroxylation site or not. Once labelled, instances are then added to the training set *L* and removed from *U*. The process is repeated until the classifier converges, the pool of unlabelled is exhausted, and/or a predefined stopping criterion is reached.

In this thesis, we propose and evaluate a number of pool-based active learning selection strategies to design an SVM classifier for N/D hydroxylation site prediction. These active learning query strategies include the *certainty-based* query which picks unlabelled instances most likely to be a positive class, the *uncertainty-based* query strategy which selects unlabelled instances closest to the classifier decision boundary (i.e. most uncertain or most informative), and the *density-uncertainty-based* query which adds careful selection of the initial training set using instances that are most representative.

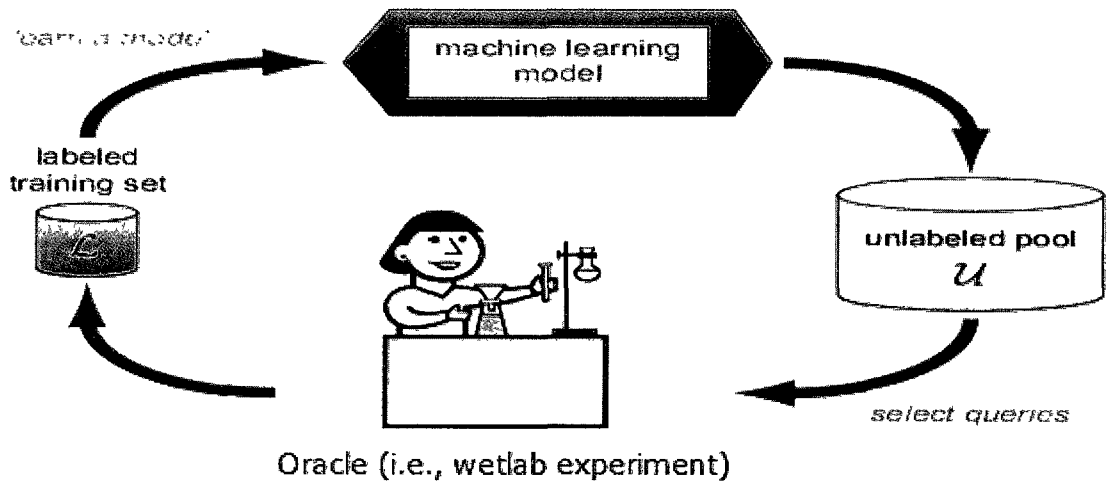

Figure 1.2. The pool-based active learning cycle (reproduced from (Settles, 2010)).

These active learning strategies are shown to effectively select, from a pool of unlabelled N/D sites, the most informative instances to be labelled by wetlab experiments in order to maximize classifier learning while requiring fewer labelled training data. The benefit of this strategy is to drastically reduce the amount of manual labour required to annotate (i.e. label) protein instances, thereby requiring fewer wetlab experiments to achieve the desired level of classifier performance.

The proposed approaches are compared with the traditional approach of passive learning, where labelled training data is collected by randomly selecting unlabelled points for manual annotation through wetlab experiments. In passive learning, the classifier works through this pool of unlabelled data, thereby risking inclusion of redundant instances, which may fail to increase the quality of the model for the prediction of N/D hydroxylation sites.

Finally, wetlab experimental validation of a number of putative N/D sites identified by the active learning query strategy has been conducted as part of this thesis. The proteins of interest have been isolated, quantified and over-expressed in order to determine the true annotation of these putative N/D hydroxylation sites.

1.2 Motivation

Protein hydroxylation site identification is often successfully tackled by machine learning methods that can automatically categorize new N/D hydroxylated sites. Prediction systems are normally sequence-based, since solved protein structures are only available for a small number of proteins. However, supervised machine learning algorithms require labelled N/D training data, from which the predictive classification model is learned. This process involves extensive wetlab experiments and is a costly procedure. Therefore, this annotation effort can be reduced significantly when data points to be annotated are carefully chosen through active learning strategies. Hence, in the current thesis, we have proposed a pool-based active learning to meet this particular requirement. As such, the most useful and relevant instances are intelligently selected from the pool of unlabelled N/D data points to be labelled through wetlab experiments, and subsequently added to the training data to learn an improved predictive model. Consequently, the same level of performance can be achieved while requiring fewer instances than with the traditional machine learning process (passive learning).

1.3 Problem Statement

The thesis explores key aspects of active learning for the prediction of N/D hydroxylation sites on human proteins. Specifically, we seek to:

i. Develop a pool-based active learning query strategy that considers classification scores among the unlabelled protein data set to determine how "informative" each unlabelled instance (i.e. N/D site) is to the learner. In this way, the most informative unlabelled points can be labelled as positive or negative via wetlab experiments in order to maximally increase the learning of the prediction system.

ii. Implement active learning algorithms that can intelligently identify putative N/D hydroxylation sites that have the highest probability of successful experimental verification and validation through wetlab experiments. This strategy is most appropriate when an experimenter is more interested in identifying novel hydroxylation sites rather than maximally improving the prediction system itself.

iii. In some cases, there is no initial labelled training set. In this case, an active learning strategy can be applied that also considers the representativeness of each instance to create the initial training set. By selectively querying the most representative and most informative instances, we can further reduce annotation effort with improved performance.

iv. Considering the interdisciplinary nature of the Masters in Biomedical Engineering Program, this thesis goes beyond strict simulation of active learning to also encompass wetlab validation of a number of putative N/D hydroxylation sites selected via active learning. This includes the bioinformatics analysis required to account for wetlab experimental considerations when selecting which proteins should be annotated through experiments, and also participating in the actual wetlab experiments including protein expression, isolation, quantification, and characterization through mass spectrometry.

1.4 Contributions

The contributions of this thesis are as follows:

Phase I: Simulation of active learning

A comprehensive comparison was performed between uncertainty-based query, density-uncertainty-based query and certainty-based query strategies in terms of prediction recall, precision, Matthews' correlation coefficient, area under the receiver operating characteristics curve, and number of labelled positive and negative instances with growing training set size. We considered the default passive learning approach as the benchmark throughout this thesis. An extensive performance comparison of these methods relative to the N/D dataset is provided. We have clearly demonstrated that the active learning cycle can drastically reduce the amount of annotation efforts required to obtain a given level of classification performance for the prediction of hydroxylation sites. To the best of our knowledge, a comprehensive comparison between various active learning query strategies for the prediction of a protein post-translational modification has not been reported in previous literature. Our results from the simulation of active learning query strategies for the prediction of N/D hydroxylation sites on human proteins have been accepted for publication in the proceedings of the International Conference on Computational Intelligence and Bioinformatics (CIB 2011), Nov 7-9 2011 in Pittsburgh PA.

Phase II: Wetlab experimental validation

Unlike other studies that only simulated the active learning process (e.g. Mohamed et al., 2010; Liu, 2004), in this thesis wetlab experimental validation/verification of putative N/D hydroxylation sites identified for labelling by applying active learning query strategies is considered and performed. Instead of validating all 1.3 million putative hydroxylation sites identified by Liu's (2009) predictive tool, we have used active learning to identify the 20 top-ranked N/D hydroxylation sites by both uncertainty-based and certainty-based sampling methods. Wetlab experimental requirements and considerations were then taken into account, through the application of bioinformatics analysis, to ultimately select the top 3 proteins identified by certainty-based sampling and 2 proteins by uncertainty-based sampling that were most suitable for wetlab validation. Our choice was influenced by the availability of the protein plasmids and antibodies, the suitability of the tryptic fragments containing the putative sites for analysis via mass spectrometry, and the associated cost of materials used in the wetlab validation. Wetlab experimental validation of 2 putative proteins identified by the active learning query strategy have been isolated, quantified, and over expressed as protein of interest in mammalian cells, human embryonic cell 293 (HEK293) that is most likely to ionize in mass spectrometry (MS) for the determination of the true annotation of these putative N/D hydroxylation sites.

1.5 Overview of Results

The overview of results can be discussed in two-stage phases:

Phase I: Simulation of Active Learning

We have successfully developed a pool-based active learning with different query strategies that considered only the most informative unlabelled instances from a pool of unlabelled data points based on the predefined classification scores.

We analyzed the viability of active learning and implemented various active learning query strategies, uncertainty-based, density-uncertainty-based, and certainty-based sampling techniques with SVM. These query techniques intelligently and successfully proved to be suitable for purpose of the research goals, exemplified by the empirical results expressed in terms of recall (sensitivity), predictive positive value (PPV) or precision rate, Correlation Coefficient (CC) and Area under the receiver operating characteristic (ROC) curve (AUC). After 50 iterations of active learning, the uncertainty-based sampling technique achieved a recall value of 96% ± 0.03, 98% ± 0.04 PPV, 0.87 ±0.03 MCC and 0.98 ±0.00AUC values; density-uncertainty-based sampling achieved a 97% ± 0.02 recall, 100% ± 0.00 PPV, 0.89 ± 0.02 MCC, and 0.98 ± 0.00 AUC; certainty-based sampling achieved a recall value of 100% ± 0.00, 80% ± 0.02 PPV, 0.83 ± 0.20 MCC and 0.98 ± 0.00 AUC; and passive learning can recall 90% ±0.03, 97% ± 0.05 PPV, 0.83 ± 0.04 MCC and AUC value of 0.96 ± 0.03 for 50 queries or iterations. Out of 35 positive instances available to query strategies, certainty-based identified all 35 instances, while both uncertainty-based and density-uncertainty-based techniques identified 21 and 22 instances respectively and passive learning identified only 4 instances. Lastly, our empirical results have shown that the uncertainty-based active learning strategy is an effective way to handle class imbalance among the available unlabelled data.

Phase II: Wetlab Experimental Validation

We have utilized uncertainty-based and certainty-based active learning query strategy to intelligently identify 20 potential N/D hydroxylated sites from 1.3 million putative hydroxylation sites that are most likely to improve classification accuracy.

We successfully obtained all 20 protein sequences from NCBI database to examine and detect the suitability of the tryptic fragments (polypeptides) containing the putative sites for analysis through mass spectrometry (MS) for the determination of the true annotation of these putative N/D hydroxylation sites.

Five (5) out of the 20 proteins were chosen for wetlab experimental validation; after bioinformatics evaluations on the availability of genes encoding these target proteins, the availability of specific antibodies for extracting the target proteins, and a simulation of the trypsin digestion process were employed to unveil N/D hydroxylated sites.

Four (4) of the target genes (TP53BP2, PPP1R13L, AP2M1 and CCBE1) were successfully isolated from the plasmid vectors of type pCMV-SPORT6 and expressed as proteins in mammalian cell lines i.e. human embryonic kidney cells 293 (HEK 293). The fifth target gene, LTBP2 was successfully subcloned from pBlueSCriptR into pCMV-SPORT6.

TP53BP2 and AP2M1 proteins were successfully isolated, quantified, over-expressed in HEK 293 cells in an *in vitro* under normoxic and hypoxic experiment. These proteins biological activity was verified using Western blotting, immunoprecipitation and Coomassie stain analysis based on the proteins expression identified on an SDS-PAGE gel. These specific bands were excised from the gel, transferred to Eppendorf tubes and stored at -80°C. The successful identification of these proteins' expression on the gel lays the foundations for the determination of the true annotation of these putative N/D hydroxylation sites via mass spectrometry. Once validated to be positive or negative, these new results will be added to the training set and the performance will be evaluated on an independent test set to complete the active learning cycle. This final stage of analysis is left to future work. However, the vast majority of wetlab experiments have been completed for these two target proteins.

1.6 Organisation of Thesis

The remainder of this thesis is organized as follows:

Chapter two presents detailed background information on protein synthesis, biology of hydroxylation, pattern classification, active learning algorithms, different active learning methods, query strategy frameworks, active learning specifically with SVM, passive learning strategy, and an active learning stopping criteria.

Chapter three describes dataset collection, feature selection, data pre-processing, and the splitting of the dataset into the initial training set, the test set, and the unlabelled data. It further discusses a SVM pool-based active learning strategy for binary classification, measures of an instance's informativeness, and proposes an active learning pool-based framework. Specific active learning

strategies to be evaluated in this thesis are defined, along with performance evaluation metrics. Experimental results achieved for each active learning system, and a comparison of the different active learning querying strategies with respect to passive (baseline) learning strategy are included. Finally, an active learning stopping criterion was looked at.

Chapter four describes a detailed implementation of active learning on a real life dataset including an experimental demonstration of active learning in a wetlab scenario. Wetlab experimental protocols and results are described for the validation and verification of a number of putative hydroxylation sites on five proteins. The bioinformatics analysis conducted to select the 5 putative hydroxylation sites among those sites highly ranked by active learning is also described.

Chapter five provides a brief summary of contributions and recommendations for future work.

2 CHAPTER: LITERATURE REVIEW

In this chapter, the literature is reviewed to provide the reader with the necessary background and context for this thesis. The review begins with protein biosynthesis describing transcription, translation and the biology of hydroxylation involving asparagines and aspartate amino acids. It then continues with the concept of pattern classification, different types of supervised and unsupervised learning methods and related work on the prediction of hydroxylation sites on proteins. Support vector machines are described, as this is the main machine learning technique to be used with active learning in this thesis. Active learning with different query strategies is then discussed. Lastly, a stopping criterion for active learning with SVM classifiers is described.

2.1 Biology of Protein Hydroxylation

2.1.1 Protein Biosynthesis (Synthesis)

Protein synthesis is the process where cells build proteins, beginning with transcription and ending with folding and post-translation modification.

Transcription

Transcription is the synthesis of RNA (ribonucleic acid) from a template (noncoding) strand of the *DNA* (deoxyribonucleic acid) double helix. Transcription involves three phase-processes: initiation, elongation and termination.

Initiation phase of transcription

Here, an *RNA* (Ribonucleic acid) *polymerase*, an enzyme, binds to a specific region on the *DNA* that designates the starting point of transcription, called *promoter*. As the *RNA polymerase* binds on to the promoter, the *DNA* strands begin to unwind.

Elongation phase of transcription

RNA polymerase continues to travel along the template (noncoding) strand, synthesizing a ribonucleotide polymer. The RNA polymerase uses the *DNA* from the noncoding strand as a template to copy the coding strand.

Termination phase of transcription

As the polymerase reaches the termination, modifications are required for the newly transcribed messenger *RNA* or *mRNA* to be able to travel to the other parts of the cell. A *5' cap* is added to the *mRNA* to protect it from degradation. A *poly-A tail* is added to the *3' end* for protection and as a template for further processing.

Translation

During translation *mRNA*, previously transcribed from *DNA*, is decoded by *ribosomes* to make proteins. The *ribosome* has sites which allow other specialized transfer *RNA (tRNA)* to bind to the *mRNA* at the start codon (AUG); here the codon means the coding of the *mRNA* sequence as a unit of three nucleotides. The binding of the correct *tRNA* to the *mRNA* on the ribosome is accomplished by an "anticodon" that is part of the *tRNA* and a "codon" on the *mRNA*. This

correct *tRNA* is chemically linked to a specific amino acid, which is then directed to the *ribosome* to be added to the growing polypeptide (Voet et al., 2009; Lodish et al., 2008).

As the *ribosome* travels down the *mRNA*, one codon at a time, another *tRNA* comes into a second *ribosome* site. A peptide bond is formed between the two amino acids and the ribosome moves one codon "down" the *mRNA*. The first *tRNA* is released, then, the amino acid and its attached tRNA moves to the second site in the ribosome, freeing up the first to accommodate another incoming amino acid-tRNA. This process continues, until a stop codon on the *mRNA* is reached and a long chain of amino acids (protein) is produced. At this point in time, the *ribosome* falls apart and the newly formed protein is released, indicating the termination point of the transcription-translation process. During and after protein synthesis, events such as post-translation modification and protein folding occur (Voet et al., 2009). Various steps involved in the protein synthesis are shown in Figures 2.1, 2.2 and the Appendix for a detailed explanation. In this thesis, we focus mainly on the post-translational modification of proteins, which is described in the following subsections.

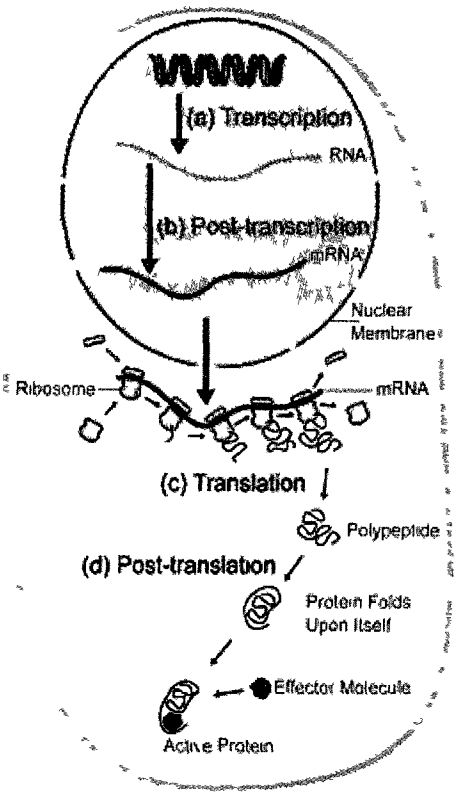

Figure 2.1. Diagrammatic representation of protein synthesis (source: http://www.ncbi.nlm.nih.gov/About/primer/genetics_cell.html).

The transcription-translation process begins at the cell's nucleus, where genes (*DNA*) are transcribed into *RNA*. Then, post-transcriptional modification processes modify the *RNA* into mature m*RNA* which is transported into the cytoplasm from the nucleus. During translation, the m*RNA* becomes translated by the ribosomes by matching the three nucleotides (i.e. codons) of the m*RNA* to the appropriate three nucleotides of the t*RNA*. Optionally, post-translational processes binding the newly synthesised protein to an effector molecule to become fully active protein. An effector molecule is a small molecule, which could be a sugar, amino acid or

nucleotide that binds to protein regulator, and therefore changes its ability to interact with operator.

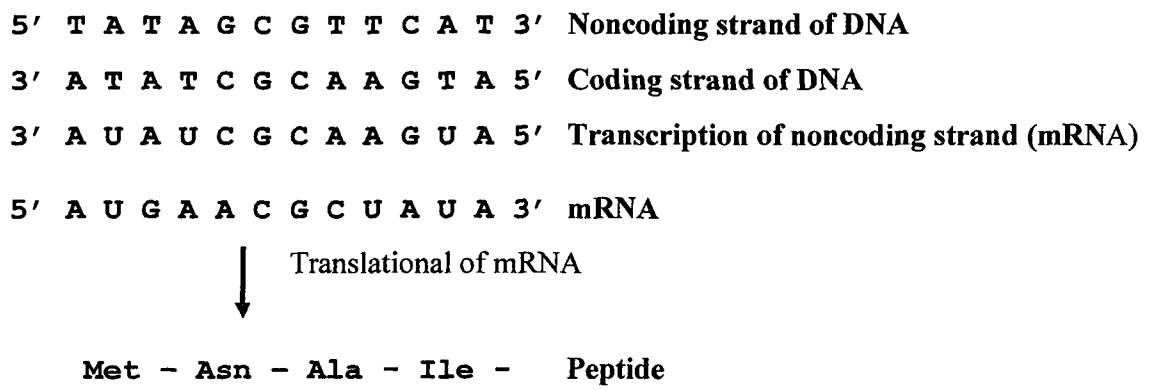

Figure 2.2. Sequence representation of protein synthesis.

DNA transfers information to *mRNA* in the form of a code defined by a sequence of nucleotides bases. During protein synthesis, *ribosomes* move along the *mRNA* molecule and "read" its sequence three nucleotides at a time (codon) from the 5' end to the 3' end. Each amino acid is specified by the *mRNA*'s codon and pairs with a sequence of three complementary nucleotides carried by a particular tRNA (anticodon). Then, after translation of mRNA an individual polypeptide is produced and can be folded and modified further by post-translational modifications. See Appendix A.2 for more details.

2.1.2 Post-translational Modifications of Proteins

Many proteins undergo some form of post-translational modification (PTM) after translation. PTM of proteins involves the attachment of biochemical functional groups, such as hydroxyl (OH), acetate, phosphate, various lipids, carbohydrates, etc., to the amino acid side chains of

proteins. Reversible modifications (e.g., phosphorylation) are involved in a variety of biological processes in humans including signal transduction and enzyme activation/inactivation; while non-reversible modifications may consequently lead to protein degradation, pathological changes and undesirable diseases (Lee, et al., 2006; Basu and Plewczynski 2010). Most commonly observed PTMs are hydroxylation, phosphorylation, and glycosylation, but this thesis focuses hydroxylation.

Hydroxylation is an important, oxygen-dependent protein post-translational modification. During the physiochemical process of hydroxylation, a protein amino acid side chain is modified by the attachment of at least one hydroxyl group (OH) (Hu et al., 2010) in a reaction catalyzed by enzymes known as hydroxylases. Hydroxylation is a reaction that is also dependent upon iron, ascorbic acid (vitamin C) and α-ketoglutarate. Hydroxylation is important for oxygen sensing in cells, and lack of protein hydroxylation signals low oxygen (hypoxic) conditions within the cells. Therefore, it is important to analyze protein hydroxylation and its relationship to human diseases (such heart diseases, cancer and diabetes) that are known to be associated with low oxygen.

2.1.3 Hypoxia

Hypoxia literally means "deficient in oxygen". Hypoxia is a reduction of oxygen (O_2) supply to a tissue below the ambient levels found in the body. Our Earth's atmosphere is composed of 21% oxygen and levels below this are considered hypoxic. Hypoxia elicits in a wide range of adaptive responses a) at the systemic level, b) at the tissue level and c) at the cellular level. Responses at the systemic level increases the alveolar ventilation thereby promoting survival by maintaining the arterial blood hemoglobin saturation and systemic O_2 transport. At the tissue level, hypoxia

stimulates the production of various growth factors including vascular endothelial growth factor (VEGF), which promotes capillary growth and sustain local tissue O_2 delivery. At the cellular level, hypoxia elicits an increase in the expression and secretion of the hormone erythropoietin, which increases systemic O_2 supply by amplifying the rate of erythrocyte formation. Most of the genes that are activated during hypoxia are regulated by a single transcription factor known as the hypoxia inducible factor (HIF), (Schumacker 2002; Giaccia et al., 2004).

2.1.4 Relationship between Hypoxia Inducible Factor (HIF) and Hydroxylation

Hypoxia is the physiologic trigger that activates hypoxia-inducible factor (HIF). HIF is a transcriptional factor that acts as a major regulator in the detection and responses to hypoxia (i.e., low oxygen level) in the human tissues (Peet & Linke 2008). The HIFs (HIF-1α and HIF-2α) are key transcription factors regulating the expression of most, but not all, hypoxia-inducible genes. The HIF is a heterodimeric protein comprised of an alpha subunit and a beta subunit. Under normal oxygen conditions (normoxia), the alpha subunit of HIF is hydroxylated on two proline residues (Pro402 and Pro564 of the human protein) by the prolyl hydroxylase domain (PHD)-containing hydroxylases (Lando et al., 2002). Hydroxylated HIF alpha is recognized and bound by the von Hippel Lindau (vHL) protein; a protein that, when mutated, leads to a clear cell carcinoma that develops in renal cells. The vHL protein is also part of a much bigger complex of proteins that form an E3 ubiquitin ligase. The E3 ubiquitin ligase subsequently links ubiquitin, a short polypeptide that signals protein degradation, to the alpha subunit of HIF. Ubiquitin is attached in long chains and this signals the cells to degrade the ubiquitinated protein. Ubiquitinated proteins are degraded by a large protease known as the proteasome. There is a cytosolic and a nuclear proteasome that degrade cytosolic and nuclear ubiquitinated proteins

respectively. Degradation of HIF alpha, from hydroxylation to proteolytic degradation, occurs with a half life of 5 minutes. Thus, it is one of the more rapidly turned over proteins known. In the absence of oxygen (hypoxia, anoxia), HIF alpha is not hydroxylated, not ubiquitinated and not degraded by the proteasome, and thus can travel to the nucleus of the cell (through the use of a nuclear localization sequence of amino acids present within the protein), bind to its partner protein (HIF beta) and bind to a DNA sequence known as the hypoxia response element (HRE) found in close proximity to most (but not all) hypoxia-inducible genes.

HIF alpha is also hydroxylated on an asparagine residue within the C-terminal transactivation domain of the protein (Asn-803) by an asparagines/aspartate hydroxylase known as Factor Inhibiting HIF (FIH) as shown in Figure 2.3. Oxygen-dependent hydroxylation of this residue prevents protein-protein interaction between HIF alpha and its coactivator of transcription, the CREB binding protein (CBP) or p300. Thus the presence of oxygen not only causes HIF alpha to be rapidly degraded, but it also disrupts the transactivation of hypoxia-inducible gene expression by the protein. FIH has also been shown to catalyze the hydroxylation of highly conserved asparagines (*N-OH*) residues within the ubiquitous ankyrin repeat domain (ARD)-containing proteins (Peet & Linke 2008; Peet et al., 2004; Cockman et al., 2009). The effects of hydroxylation appear to be predominantly localized to the target asparagine and proximal residues, at least in the consensus ARD-containing proteins (Hardy et al., 2009).

Figure 2.3. The crystal structure of factor-inhibiting hypoxia-inducible factor (FIH) reveals the mechanism of hydroxylation of HIF-1 alpha (Elkins et al., 2003).

FIH (in *red*), asparagine (Asn-803) in *green*, $Fe^{(II)}$ and 2-oxoglutarate are show as *ball-and-stick* representation in *grey* and alpha helix in *cyan*. The FIH is an asparaginyl β-hydroxylase and is a member of 2-oxoglutarate and $Fe^{(II)}$ –dependent dioxygenases which catalyze the hydroxylation of Asn-803 of HIF (Lancaster et al., 2004).

2.2 Pattern Classification

2.2.1 Introductory Concept of Pattern Classification

Classification tasks occur in a wide range of human activity. Classification that involves devising a decision rule to be applied repeatedly in order to classify data points into one of a set of pre-defined classes based on features is termed pattern classification. In supervised learning, a

training dataset with known associated class labels is used to determine a function or decision rule (i.e., set of decision boundary) that accurately maps the features to the true class labels. The mapping function can be expressed explicitly in the models (classifiers) or implicitly in the data (Duda et al., 2001). A typical pattern classification problem is illustrated in Figure 2.4, where the goal would be to determine a rule that would differentiate blue points from green points. Generally, the learning methods used for pattern classification problems fall into two categories: supervised learning and unsupervised learning methods.

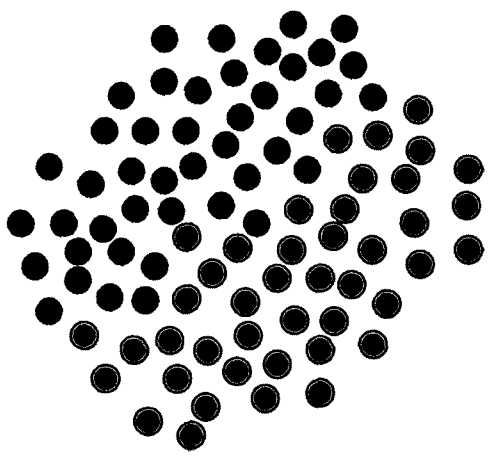

Figure 2.4. Example of classification pattern in a 2D input space.

Class labels are indicated by point colouring.

2.2.2 Related Terminology

(i) **Instances:** An instance x represents a specific object. The instance is usually represented by a d-dimensional feature vector $x = x_1, \ldots, x_d \in \mathcal{R}^d$, where the length of the feature vector, d, is known as the dimensionality of the feature vector. The x represents the whole instance, and x_d denotes the d^{th} feature of x. In this thesis, each N/D site is an instance and is represented by a feature vector of length 360.

(ii) **Label:** A label y is the desired prediction on an instance x. Labels form finite set of values; these distinct values are called classes, usually represented by integer numbers {-1, 1} for a binary classification problem, thus $y \epsilon \{-1, 1\}$. This is regarded as binary class labels, and the two classes are generically called the negative class (non-hydroxylated site) and positive class (hydroxylated site) respectively.

(iii) **Training Sample:** A training data set is a collection of instances $\{x_i\}_{i=1}^{n} = \{x_1, \ldots, x_n\}$, which acts as the input to the active learning process. In a supervised learning, the training sample consists of pairs, each containing an instance x and a label y: $\{(x_i, y_i)\}_{i=1}^{n}$. Thus, y is a label for x provided by a teacher or supervisor, hence the name *supervised* learning. Therefore, such (instance (x), label(y)) pairs are called *labelled data,* while instances without labels are called *unlabelled data.*

2.2.3 Supervised Learning Algorithm

Supervised machine learning model is a branch of pattern classification, which automatically induces a predictive model from labelled data. Thus, given a collection of data points, a supervised learning model predicts the label associated with any new data point based on a set of observable features that describes it. In general, supervised learning classification can be characterized as follows. The learner is provided with a labelled set of training instances, $L = \{(x,y)_1,\ldots,(x,y)_n\}$, with which to induce a model. A model means a formal representation with an interpretation, under which instance inputs are mapped to label outputs. The model defines a classification function f on the instances: $y = f(x)$. Therefore, the main objective of the training procedure is to find the "best" or optimal model according to some *objective function* for a given problem within the defined *model space* (or *hypothesis space*) (Settles, 2008; Duda et al., 2001).

2.2.4 Types of Supervised Learning Methods

Some of the supervised learning methods for pattern classification are briefly described below to attest for the choice of our current model. The supervised learning methods include support vector machines, decision trees, K-nearest neighbour, artificial neural networks and Naïve Bayes.

Support Vector Machines

Support vector machines (SVMs) are a form of supervised machine learning based on statistical learning theory (Vapnik, 1999) for solving classification problems and have achieved excellent performance on a wide variety of bioinformatics classification tasks. SVMs can implicitly map data from the original low-dimensional space into a high-dimensional space through application of kernel functions and learns a decision boundary (i.e., hyperplane) which eventually separates the training data points into two different classes (Burges, 1999). The hyperplane in the higher–dimensional space is selected based on the maximal margin between the two classes. The goal is to produce a classifier that will generalize well on unseen data points. SVM are equipped with structural risk minimization which enable SVM to generalize well (Vapnik, 1999), which differentiated it from other supervised machine learning approaches for many pattern classification problems. SVMs are used throughout this thesis and are described in detail in section 2.3.

Decision Trees Induction

A decision tree is a non-probabilistic supervised learning approach, which relies on classification rules and uses the tree structure to classify instances. The decision tree classification starts from

the root of the tree, and one or more feature(s) of the instance is compared to a specified function to determine the branch to follow. Another feature will be compared to a new specified function in the next internal round. Therefore, the comparison continues until the said instance reaches a leaf node, and associated class label is assigned to the instance. The decision tree learning system can often produce performance values considerably inferior to other supervised learning models due to algorithm instability and class-overlap problem (Quinlan, 1993).

Naïve Bayes Model

Naïve Bayes is a probability-based method. It models the joint probability $P(x, y)$ of the labelled training instances, and uses Bayes' rule to predict label probabilities (Hand & Yu., 2010). Naïve Bayes classifiers work under the assumption of conditional independence which states that features are independent from each other given knowledge of the output class. Additionally, naïve Bayes is a generative model for which training relies on the estimation of the likelihood $p(x \mid y)$ from the training data (Friedman et al., 1997). This estimation is inaccurate in the case of active learning since the training data are not randomly collected (Nguyen & Smeulders, 2004; Guoliang, 2009).

K-Nearest Neighbour

Neighbour K-nearest neighbour (KNN) classification model maintains no model parameters but assigns labels to an instance based on similarity of instances in the training set. The mapping function from features to the class labels is implicitly expressed in the training set (Aha, et al., 1991). A KNN learning system prediction time could be longer as it searches for similarity between the previous instances and each new instance before it makes a prediction.

Artificial Neural Network Model

Artificial neural network is a method inspired by a biological neural system which consists of many neurons. The neurons in artificial neural network are interconnected and work together to realize a mapping function. The links between neurons can be trained with a data to strengthen the particular patterns. The representative training method for artificial neural networks is back-propagation. A neural network can approximate any functions accurately, provided the number of the neurons, connection function, and weights of the connections are properly selected. Traditional neural network approaches often have generalization, and data overfit problem as a result of the optimization algorithms used for the parameter selection (Guoliang, 2009; Bishop, 1995).

2.2.5 Unsupervised Learning Method

In an unsupervised learning algorithm, there is no supervision of how individual instances should be handled. One of the common tasks of unsupervised learning is clustering, where the main objective is to separate the n instances into clusters; similar instances are placed in the same cluster/group, and instances in different clusters are dissimilar. One of the common clustering algorithms under this setting is the K-means clustering algorithm. The K-means is also used in density-uncertainty-based query strategy to create an initial training set in this thesis, and therefore the principle and functionality of the K-means algorithm is discussed with illustrative flowchart representation below.

The K-means clustering algorithm uses an iterative refinement technique to partition the instances into K clusters, where K is a user-defined number of clusters. The algorithm begins by

randomly determining initial cluster centroids (centres) of each K clusters (i.e., randomly picks data points from the dataset to create initial cluster centres). A centroid is an artificial point in the input space of the instance which represents an average location of the particular cluster. Then the distance between each instance and its cluster centroid is calculated to reconstruct a new partition by associating each instance to its closest centroid. Subsequently, the centroid of every newly created cluster is recalculated; and the algorithm is repeated until the instances no longer switch clusters or the centroids no longer change (Duan & Babu, 2008). Optionally, once the K-means algorithm converges to the final clusters, a representative instance may be selected from each cluster to represent the final clusters. It is assumed in all clustering algorithms that distances between instances are representative i.e. instances close to each other in feature space are also similar to each other in reality. A diagrammatic representation of K-means is depicted in Figure 2.5.

2.2.6 Related Work on Hydroxylation predictions

(Liu, 2009) demonstrated the prediction of hydroxylation sites based on dataset taken from experimentally verified and validated N/D hydroxylation sites of 40 proteins known to have at least one N/D hydroxylation sites. The input features consisted of surface accessibility, secondary structure, and Position-Specific Scoring Matrix (PSSM) data and detail information of these features is discussed in Section 3.2. All feature vectors make up the dataset containing 1813 non-identical data samples, with 55 positive and 1758 negative data points, were used to induce SVM-based prediction model.

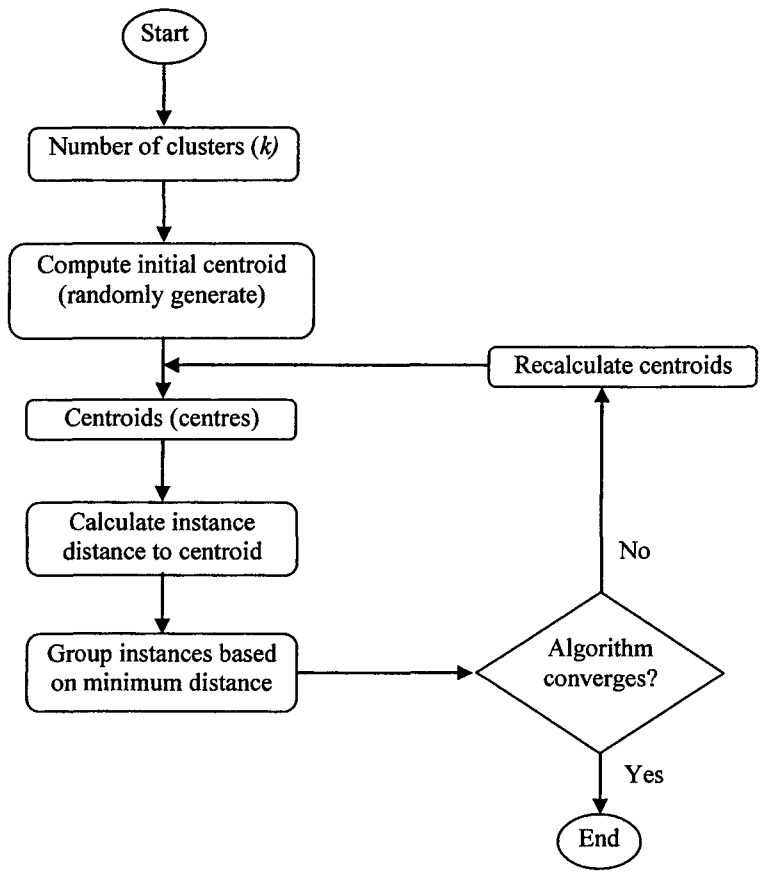

Figure 2.5. Flowchart representation of K-means algorithms (reproduced from Khan and Mohamudally, 2010).

This illustrates the two basic steps involved in the implementation of K-means algorithm. The first step is the assignment step where instances are placed in the closest cluster. The second step is the re-assignment step where the cluster centroids are recalculated from the instances within each cluster.

This class imbalance could undermine the useful information provided by the positive class; as such, the prediction model would be biased towards the dominant negative class. This is a typical characteristic of all traditional supervised learning algorithms when faced with large class imbalance in the training data. Nevertheless, Liu addressed this class imbalance by random undersampling without replacement using a subset of the negative (dominant) class to train the SVM-based classifier. The effectiveness of the random undersampling approach was measured by training the classifier on the ratio of positive/negative data points as 1:1, 1:2, 1:3, etc and 1:1, 2:1, 3:1, 4:1, and validated by leave-one-out test method. Utilizing this approach a recall of 92.73%, 74.3% Matthew's correlation coefficient (CC) and positive predictive value (precision rate) of 61.45% was achieved, using a positive: negative training ratio of 1:3. Furthermore, this study revealed that all positive N/D hydroxylation sites are located on the protein surface and occurred in non-regular or beta-strand secondary structures. This study also identified 1,288,896 potential N/D hydroxylation sites among all human proteins.

It would be resource-intensive and time-consuming to experimentally verify all 1.3 million potential asparagine and aspartate hydroxylation sites obtained by Liu's (2009) predictive model, hence the motivation for the current study. Therefore, this study aimed at developing a supervised active learning computational model to determine which putative hydroxylation sites should be selected for characterization via wetlab experimental validation and verification. By further improving the accuracy of the prediction model, we will help gain useful insights and understanding of the complex physiochemical mechanism of hydroxylation of N/D on human proteins.

In another study, Hu et al. (2010) developed a prediction tool to identify sites of proline and lysine hydroxylation. They established that there is a relationship between the three kinds of amino acid features for protein hydroxylation sites : i) amino acid indices (AAindex), represents the physiological and biochemical properties of the amino acids, ii) Position-Specific Scoring Matrix (PSSM) which represents evolution information of amino acids, and iii) structural disorder of amino acids. This supervised prediction model for identifying proline and lysine hydroxylation sites utilized a nearest neighbour approach. Evaluation through jackknife cross-validation obtained empirical results for sensitivity, specificity and Matthew's correlation coefficient of 64.8%, 81.6%, and 0.461 respectively utilizing hydroxyproline dataset, and 70.4%, 88.0%, and 0.592 respectively with hydroxylysine datasets. They observed that the physiochemical, biochemical, and evolution information play significant roles in the identification of protein hydroxylation sites, while structural disorder had a less significant role in the hydroxylation of proteins.

2.3 SVM Formulations

The SVM formulation, let data points $\{x_1,\ldots,x_n\}$ be the vectors in the feature space $X \in R^p$ with associated labels $y_i \in \{-1,1\}$. The hyperplane is given by $(w \cdot x) + b = 0, w \in R^p$, b $\in$ R. Given a linearly separable case, the SVM classifier is able to generate an optimal separating hyperplane with maximum generalization ability (i.e., it uses vector w and parameter b to minimize $\|w\|^2$) such that each instance is classified as positive or negative i.e.,$(w \cdot x_i + b) \geq 1$, if $y_i = 1$; $(w \cdot x_i + b) \leq 1$, if $y_i = -1$. This can be rewritten as $y_i\,[(w \cdot x) + b] \geq 1$ for all $1 \leq i \leq 1$. This is the quadratic program (Zhan & Shen, 2004; Byvatov, 2011). The w is a weight vector

normal to the hyperplane; |b|/||w|| is the perpendicular distance from the hyperplane to the origin (Liu 2004).

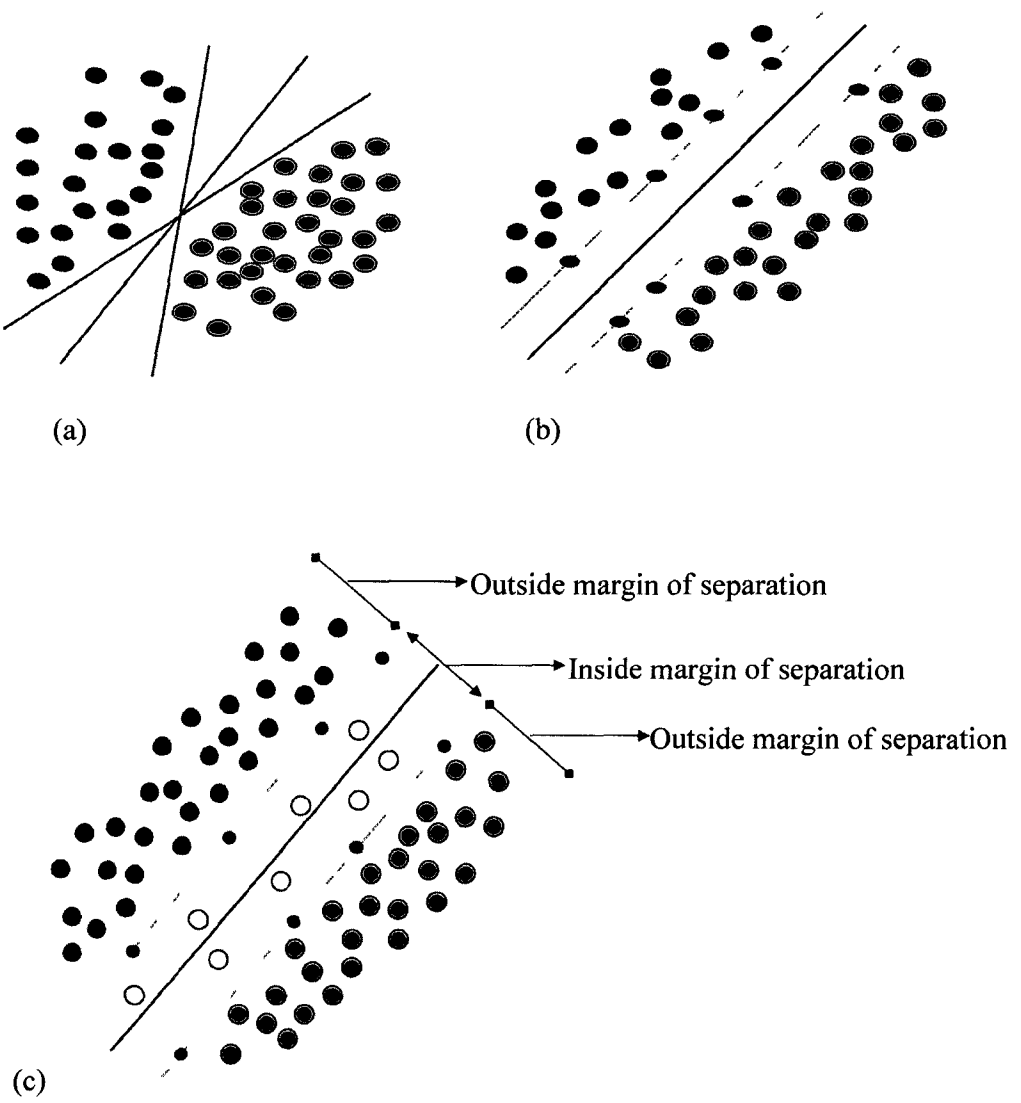

Figure 2.6. (a) Schematic representation of three possible separating hyperplanes. (b) Schematic representation of an optimal separating hyperplane. (c) An optimal separating hyperplane with unlabelled test points inside the margin. Figure 2.6(b) and (c) shows the separating hyperplane (solid lines), margins (dashed lines) and the support vectors of the SVM (black circles). The positive and negative training

instances are indicated by purple circles and green circles and unlabelled test points represented by open circles.

The margin is the distance between the two boundary lines parallel to the hyperplane. All training instances lying on the margins are called support vectors indicated as the black circles shown in Figure 2.7 (b) and others laying within the margin as shown in Figure 2.7 (c). This idea introduces us to the concept of hard-margin and soft-margin SVM classifiers. In the hard-margin SVM classifier, there is no support vectors permitted to lay within the margin and the SVM is trained to maximize the margin (i.e., the distance between the separating hyperplane and nearest training instance) (Liu, 2004). Whereas in the soft-margin SVM classifier (Tax et al., 1997), a trade off is introduced between maximizing the margin and minimizing number of observed training errors. Here, some of the support vectors lying within the margin may be classified incorrectly by the separating hyperplane, which is regarded as the trade off between having a maximized (large) margin and a minimized number of errors on the training set. This trade off is introduced by C parameter, which must be optimized to achieve maximum classification accuracy (Cortes & Vapnik, 1995; Byvatov, 2011).

For training instances that cannot be linearly separated in the original input feature space, the SVM makes a non-linear transformation of the original input vectors into a potentially higher dimensional *feature space,* where an optimal separating hyperplane can be found. The margin is maximized within this feature space to enhance the generalization ability of the classifier (Gunn, 1998). The unique solution can be achieved by using kernel functions that satisfy Mercer's condition such as polynomial, linear, radial basis function (RBF), and sigmoid (Burges, 1998).

The basic idea of the Mercer's condition is that vectors x in a finite dimension space (i.e., input space) can be mapped to a higher dimensional Hilbert space H provided with a dot product through a nonlinear transformation $\varphi(\cdot)$. Most of the transformations $\varphi(\cdot)$ are unknown, but the dot product of the corresponding spaces can be expressed as a function of the vectors:

$$\varphi(x_i) \cdot \varphi(x_j) = K(x_i, x_j) \tag{2.1}$$

These spaces are called Reproducing Kernel Hilbert Spaces (RKHS), and their dot products $K(x_i, x_j)$ are called the Mercer kernels. The Mercer theorem gives the condition that a kernel $K(x_i, x_j)$ must satisfy in order to be the dot of a Hilbert space. Assuming that there exists a kernel function $\varphi: X \rightarrow H$ and a dot product $K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)$ (Martinez-Ramon & Christodoulou, 2006; Christianini & Shawe-Taylor, 2000) then the kernel function leads to the reformulation of SVM classifier into

$$f(x, \alpha) = \sum_i y_i \alpha_i K(x, x_i) + b. \tag{2.2}$$

The kernel function is very significant, because it removes the complex computation involved in the feature space, then training the SVM classifier only requires inner products between support vectors and the vectors (data points) of the feature space. This becomes the SVM decision boundary, where α_i and b are parameters determined by the SVM's learning algorithm, while the sign of $f(x, \alpha)$ gives the predicted label of instance x (Xu et al., 2003; Zhan et al., 2005; Cui et al., 2009).

Sometimes, the training set may not be linearly separable even though it has been projected into a higher dimensional space with the application of kernel function. For such scenario, the slack penalty coefficient C is introduced to allow data points to exist on the wrong side of the separating hyperplane (i.e., concept of soft-margin discussed above). Since the value of C can vary significantly depending on the dataset, a parameter sweep using cross-validation over the training dataset may be applied to identify the best C value. The identification process will enhance the SVM classifier's generalization ability for unknown data point and controls the misclassification error during the testing phase of the classification process (Gold & Sollich, 2002; Chapelle & Vapnik, 2000). If a large value is used for C, misclassification is suppressed, and if small values are set, training instances are permitted to be misclassified.

2.4 Active Learning

Active learning is a machine learning strategy which reduces the need for manual annotation of instances for the training of supervised classifiers to attain a given performance level. For any typical supervised learning algorithm to achieve the expected performance level, it must often be trained on large number of labelled instances which represents a large investment in annotation efforts, particularly when annotation of an instance requires wetlab experiments. Furthermore, one risks inclusion of redundant training instances thereby wasting annotation effort without improving prediction performance. However, active learning aims to label the most useful or informative instances to maximize the performance of the classifier while minimizing the annotation effort of the Oracle (i.e. the wetlab experimentalist in our case). There are several active learning scenarios, and query strategies used to determine most useful (informative) instances for the active learning process. Although, this thesis focuses on *pool-based* active

learning to construct querying functions, there are other interesting active learning scenarios including *stream-based* and *membership query* which are briefly discuss in subsequent sections.

2.4.1 Active Learning Scenarios

There are three main settings of active learning scenarios, i) pool-based sampling, ii) membership query synthesis, and iii) stream-based sampling considered in literature in which active learners may pose queries to an Oracle (Baram et al., 2004; Freun et al., 1997; Cohn et al., 1994). The most common learning scenarios are pool-based sampling and stream-based sampling. This thesis focuses on the pool-based sampling learning scenario because of its applicability to the current problem. We devised techniques for querying or selecting the most informative instances from the unlabelled instances and examined if newly acquired labelled instance improves the classifier's prediction ability for the protein dataset.

2.4.2 Pool-based sampling

Pool-based active learning follows a two-stage procedure. During the initialization stage, a small number of labelled training instances, L, and a large number of unlabelled instances, U are collected, and the initial classifier is trained. The procedure then enters the iterative query stage (i.e., sampling phase) where the most informative samples are identified, labelled by the Oracle, then the classifier is retrained. Since queries of instances from the pool U are performed in a greedy fashion, an informative measure is used to evaluate all instances within the pool (Shen et al. 2004). Therefore, the most daunting task in this pool-based active learning is how to determine the "*most informative*" instances for Oracle annotation at each active learning cycle. For this task, we adopted uncertainty-based query strategies active learning (Lewis & Gale, 1994; Lewis & Catlett, 1994; Zhu et al., 2009; Xu et al., 2009; Cui et al., 2009). Here, an

unlabelled instance U with maximum uncertainty is viewed as the most informative instance; because it is the closest to the decision boundary and the current classifier (i.e., learner) is least confident about the true annotation of this unlabelled instance (Mohamed *et al.*, 2010). Unlike the strategies discussed above, the certainty-based query strategy selects unlabelled instances farthest from the decision boundary on the positive side for annotation by the Oracle.

Liu (2004) applied active learning uncertainty-based query strategy with SVM to sample from gene expression profiles of colon, lung and prostate cancers. The empirical results showed active learning significantly reduced annotation by 82% and attained area under the receiver operating characteristics curves (AUC) values of 0.81 compared to 0.5 AUC value achieved by passive learning, given the same number of labelled instances.

Doyle et al. (2009) presented and analyzed the class balanced active learning (CBAL) framework that accurately detects cancerous regions on prostate histopathology data samples. This query strategy uniquely chooses equal numbers of instances from both classes (i.e., cancer and non-cancer) to induce the active learner. The empirical results showed remarkable improvement in terms of accuracy and AUC (i.e. sensitivity and specificity) when compared with passive learning and other methods that did not specifically address the class imbalance. The CBAL required only 50 instances to achieve comparable accuracy to the full training set of 12,000 image regions used by the other methods.

2.4.3 Other Active learning Scenarios

Stream-based active learning

In this learning scenario, the learner is provided with a stream of unlabelled instances. At each iteration, a new unlabelled instance is selected and given to the Oracle, which has to decide whether to request its label or not (Baram et al., 2004; Freund et al., 1997). The distinctive difference between stream-based and pool-based active learning is that the stream-based scenario scans through the data samples sequentially and makes query decision individually, while the pool-based scenario evaluates and ranks the entire set of unlabelled data points, U, at each iteration to choose the most informative query.

Membership Query Synthesis active learning

In this active learning scenario, the active learner could query any unlabelled instance in the input space, including queries that the learner generates *de novo*, rather than being restricted to the set of unlabelled instances, U, drawn from underlying natural distribution (Angluin, 2001). Membership query synthesis has been successfully employed through "robot scientist" (King et al., 2004) to execute series of autonomous biological experiments to discover metabolic pathways in the yeast *Saccharomyces cerevisiae*. In this case, an instance is a mixture of chemical solutions that constitute a growth medium, and particular yeast mutant. Then, a label is determined by the yeast mutant thriving in the growth medium or not. These particular experiments were autonomously synthesized using active learning strategy based on logic programming, and physical laboratory robot. Performance achieved indicates active learning has led to three-fold decrease in the cost of experimental materials (King et al., 2009). However, if

the labelling is done by a human Oracle, the membership query synthesis may result in an unexpected problem. For example, when Lang and Baum (1992) trained a neural network to classify handwritten characters using membership query, many of the query images generated by the learner were unrecognizable symbols (Settles, 2010).

2.5 Active Learning Query Strategies

The effectiveness of an active learning procedure depends on the query strategy's ability to identify the most informative instances for annotation at each learning cycle. We considered uncertainty based sampling (Lewis & Gale, 1994; Tong & Kong, 2001) and density-uncertainty-based (Mohamed et al., 2010; Xu et al., 2007) query strategies with SVM to query the most uncertain unlabelled instance x_t for annotation by the Oracle at each learning cycle. We also considered the certainty-based query strategy which selects the unlabelled instance x_t farthest from the hyperplane on the positive side for annotation by the Oracle at each learning cycle. This latter approach mimics a typical experimenter's tendency to only perform follow-up experiments on the most certain predictions (i.e., those predictions which are most likely to lead to the validation of a novel hydroxylation site). In the following sections, we will provide an overview of the active learning query strategies considered in the present study.

2.5.1 Uncertainty Query Strategy

In uncertainty based sampling, instance selection is based on the uncertainty of the current classifier in its prediction. Uncertainty based sampling is model-independent and can be combined with any classifier that returns confidence or probability estimates for its predictions (Tomanek, 2010).

At each learning step, uncertainty-based query strategy interactively queries the most uncertain instance from the unlabelled pool for annotation by the Oracle. It has been previously suggested that the classifier would be most improved when selecting an instance with maximum uncertainty (i.e., the instance, for which classifier is least confident about its true classification); since this is the most informative instance (Lewis and Catlett, 1994; Mohamed et al., 2010). When using the uncertainty-based query strategy with SVM, the most uncertain instance is the instance that falls closest to the optimal separating hyperplane. Furthermore, labelling and adding such an instance to the training set is most likely to alter the SVM decision boundary (Xu et al., 2009). If these instances are selected for the next active learning cycle, this strategy has been shown to improve classifier performance (Cohn et al., 1995; Tong & Koller, 2001; Schohn & Cohn, 2000; Campbell et al., 2000).

In Tong and Koller (2001) active learning with SVM used an active learning querying strategy with the notion of version space. The version space is the set of all hypotheses (i.e. set of all hyperplanes) consistent with the training set (Mitchell, 1982; Baram et al., 2004). They suggested that using querying function to select unlabelled instance closest to the optimal separating hyperplane (decision boundary) will enhance the reduction of the version space. This querying function is called "Simple Margin". To overcome the inherent problem in the aforementioned querying function, they used refined methods called "MaxMin Margin". In the MaxMin query algorithm, the authors estimated the relative size of the version space v^- and v^+ by labelling an unlabelled instance as -1 and 1, then compute margins (m^- and m^+) of the resulting SVM as the approximation of the area of the version space. Then, choose to query

the unlabelled instance for which the quantity min (m^- , m^+) is greatest. In this way, the version space is minimized and reduced for every unlabelled instances selected for labelling. This approach minimizes the error on the training set which ultimately contributed to better classification performance and reduction in the number of labels required for annotation.

The work of Campbell et al. (2000) was motivated by the SVM characteristic of constructing the decision boundary (hyperplane) using only those points which are support vectors (see Section 2.3 above). They suggested using active learning to select only those unlabelled points that will become support vectors and ignore non-support vectors, especially if one knew which instances were support vectors (Campbell et al., 2000). They opted for active learning query strategy that selects data points closest to the decision boundary. In their active learning empirical study, they achieved the best results with *sparse* data sets. These are data sets which require only few support vectors. By contrast, *dense* data sets require a relative large number of support vectors to accurately represent the hypothesis.

Schohn and Cohn (2000) applied active learning to choose a subset of the unlabelled data, and suggested that choosing a small subset serves to maximize the classification accuracy. The authors used active learning selective sampling based on SVM to select unlabelled instances that lie closest to the hyperplane, thereby maximizing the SVM margin. The empirical results showed that SVM-based active learning trained on a small subset provided a better generalization performance and required fewer data points than passive learner trained on all available datasets.

2.5.2 Representative Sampling Strategy

Representative sampling strategy queries instances that are representative of the data distribution of the unlabelled pool; because instances with highest representativeness will add more information to the training set. Here, the unlabelled instances are clustered into K groups by a clustering algorithm, and one selects the instances closest to the K centroids from the unlabelled data as the most representative and most informative instances for labelling.

Xu et al. (2003) applied representative sampling strategy for active learning with SVM, where unlabelled instances within the SVM margin are first clustered in order to identify the representative instances. These instances are taken to be the most representative and most important instances for labelling. The empirical results showed that representative sampling outperformed SVM active learning (uncertainty sampling) and random sampling during early iterations of the active learning process, however, as the training progresses its performance decreases at the end of the active learning cycles. The authors argued its poor performance could be ascribed to the poor clustering structure and high complexity of the unlabelled data within the margin.

2.5.3 Density-uncertainty Query Strategy

The density-uncertainty-based query strategy is an extension of the strategy described above, where formation of the initial labelled starting pool L is also guided by active learning. Specifically, representative sampling is used to select the instances used to create the initial training set (Hu et al., 2010; Xu et al., 2007). The density-uncertainty query strategy with SVM begins by training on this small initial training set to learn the initial classifier (Nguyen &

Smeulders, 2004) and thereafter follows the uncertainty-based sampling query strategy outlined above. A better initial training set that reflects the distribution of the data samples can improve active learning performance (Yuan et al., 2011; Hu et al., 2010; Zhu, 2008; Zhu et al., 2008; Kang et al., 2004).

Mohamed at al. (2010) demonstrated the usefulness of uncertainty sampling strategies with random and density seed active learning algorithms in the selection of protein-protein interactions to induce random forest classifiers. The data points with maximum confusion, maximum disagreement amongst decision trees in the forest, and most representative are considered as the most informative and most representative instances to train random forest classifiers for all active learning algorithms. The confusion measure was obtained as the sum of relative entropy between the average prediction values and the individual classifier predictions, among all the decision trees in the forest. They observed that all active learning algorithms used to train random forests actually achieved higher accuracy (as measured by F-score which is the harmonic mean of precision and sensitivity/recall) while requiring fewer labelled protein-protein interaction training data than passive learning strategy. In other words, they achieved 15% increase in the F-score of predicted protein-protein interactions with active learning algorithms in comparison to random selection of training data.

Hu et al.'s (2010) work reaffirmed the usefulness of the clustering algorithm to populate the initial training set to improve the performance of the active learning process. They further emphasized that using deterministic clustering algorithms, such as agglomerative hierarchical clustering (AHC) and affinity propagation clustering (APC), showed a comparable labelling

accuracy to that achieved using the best non-deterministic clustering (k-means, k-medoid). Lastly, they showed that the learning curves for active learning techniques that employed clustering techniques to create the initial training set tend to be superior to those achieved when the initial training set is selected randomly.

Nguyen and Smeulders (2004) proposed a formal model for incorporation clustering into the active learning process. They used the k-medoids algorithm to select the most representative examples to create the initial training set for their active learning model (linear logistic regression). As above, their empirical results showed that the active learning process seeded with an initial training set populated with the clustering technique outperformed other active learning algorithms such as close-to-boundary (uncertainty) and representative sampling with a randomly seeded initial training set.

2.5.4 Certainty-based Query Strategy

In this active learning query technique, the active learner queries unlabelled instances that are most likely to have positive class membership for annotation by the Oracle (Warmuth et al., 2002; Forman, 2002). In the context of SVM classifiers, at each learning step in Figure 1.1 above, this strategy selects the instance whose probability is closest to 1.0. The underlying idea of certainty-based query with SVM is that it selects unlabelled instances strongly predicted to be positive. This is a reasonable strategy when the experimenter is most interested in identifying novel positive instances rather than increasing the effectiveness of the final classifier produced (Lewis & Gale, 1994). It also emulates the experimenter's tendency to select those predictions for wetlab validation that are most likely to lead to novel positive discoveries. In this way, it

serves as a second benchmark (in addition to the passive approach) with which to compare the uncertainty-based approaches.

Forman (2002) discussed the use of supervised machine learning to predict the organic compound most likely to be active in a given binding site in order to help guide chemists' wetlab experiments to reduce cost and improve their yield. He used selection strategy termed "incremental retrain and prediction", where an instance strongly predicted to be positive is obtained from the unlabelled samples, and added to the training set of the classification process on an incremental basis to further improve the classifier's precision before the next best predicted instance (organic compound) is selected.

2.5.5 Other Active Learning Query Strategy: Query-by-Committee

In the Query-by-Committee (QBC) active learning query strategy, a diverse committee of classifiers is created from a small number of instances. Next, each committee member attempts to label additional instances, and is allowed to vote on the labelling of the instance. The instance whose annotation results in most disagreement amongst the committee members is deemed to be the most informative and is selected for annotation prior to retraining the classifier (Seung et al., 1992).

The basic idea about the QBC query strategy is to minimize errors on the training set, while maximizing the margin, which in turn translates to maximizing the generalization ability of the classifier. Since the goal of active learning is to select as few unlabelled instances for annotation as possible to achieve required accuracy, therefore QBC queries the controversial regions of the input space with a search that is as precise as possible. The implementation of QBC selection

algorithm requires construction of a committee of models that represents different regions of the classifier margin, and has some measure of disagreement among committee members (Settles, 2010).

2.6 Passive Learning Strategy

In a traditional passive learning strategy, unlabelled instances are selected for annotation at random from the large pool of unlabelled instances, *U*. The annotator (learner) has to work through this pool sample by sample; thereby risking annotation of redundant samples, which may not be helpful for learning the model. Consequently, this random selection strategy could require a large number of annotated samples to produce a classifier that meets performance requirements and may miss relevant instances that would have increased the quality of the model further. The passive learning strategy will serve as a benchmark for performance comparisons with the three active learning strategies explored in this thesis.

2.7 Stopping Criteria in Active Learning

Active learning is an interactive learning process where the active learner always interacts with the Oracle. It makes no sense, however, to continue learning until all unlabelled instances have been labelled (Zhu et al., 2008). How then, does the process know when to stop labelling instances and learning? Clearly, to minimize resources spent on the labelling of data, we wish to stop the active learning process if the selection of most informative unlabelled instances no longer contributes effectively to increasing the performance of the classifier. Lewis and Gale (1994) suggested that active learning should be stopped when the classifier has reached its maximum effectiveness. It is difficult to measure classifier effectiveness, but performance over a

hold-out test could be used as an estimate. Chen et al. (2006) further suggested that active learning should be stopped when the training set has reached a desirable size. It is realistically impossible to predefine a desirable size of training set guaranteed to improve classifier's performance effectively. Schohn and Cohn (2000) later suggested that an active learning process should be stopped when there are no unlabelled instances closer to the decision boundary within the classifier's margin.

Zhu et al. (2008) proposed a minimum expected error strategy. This stopping criteria strategy involves determining when the maximum effectiveness of the classifier is reached and the classifier's expected errors on future unlabelled instance is minimum, then, the active learner should be stopped querying the Oracle for more data points. The intuitiveness of their statistical approach is that the classifier reaches maximum effectiveness when it results in the lowest expected error on the remaining unlabelled instances.

Olsson and Tomanek (2009) used the notion of an intrinsic stopping criteria (ISC) measure of active learner stability. It involves stopping the annotation of unlabelled data when the active learner cannot learn (much) more from the data. Their concept of an intrinsic stopping criterion is based on the notion of selection agreement for query-by-committee active learning technique, similar to the approach of Seung et al. (1992) described above. The selection agreement is the agreement among the members of a decision committee regarding the classification of the most informative instance selected from the pool of unlabelled data in each round of active learning process. They suggested active learning process should be aborted when the members of the

committee are in complete agreement on instances selected from the remaining set of unlabelled data, since it is no longer expected to contribute meaningfully to the overall learning process.

Vlachos (2008) used classifier confidence estimation of unlabelled data samples to stop active learning. Specifically, active learning is stopped when the confidence estimation of the classifier follows a risk-peak-drop pattern on consecutive iterations using uncertainty-based sampling. This pattern was characterized by a rise at the beginning, then reaching its maximum value, after which it constantly drops.

2.8 Summary

In this chapter, we have described the biology of hydroxylation. We have also described various pattern classification techniques and methods on which this thesis is primarily built upon. Detailed discussion is provided of how active learning technique can reduce the amount of labelled training data required to train a classifier. A number of query functions for selecting the most informative instance to be labelled by an Oracle are described. The support vector machine is discussed as one of the best machine learning methods due to its advantages over other machine learning techniques, especially its ability to produce probability estimates and decision values well-suited for the active learning process. We also discussed active learning stopping criteria, where the iterative learning process should be stopped when there are no more useful or informative instances remaining in the pool of unlabelled data.

3 CHAPTER: EXPERIMENTAL SETUP AND ACTIVE LEARNING SIMULATIONS

3.1 Introduction

To evaluate the potential for pool-based active learning strategies in the context of asparagine/aspartate hydroxylation site prediction in proteins, we implemented the various query strategies described in Chapter 2 and ran several simulations using labelled hydroxylation site data as detailed below. All experiments began by drawing a static test set, and initial training set from the labelled N/D hydroxylation site data. The remaining data were taken as the pool of unlabelled asparagine and aspartate data from which instances to be labelled are iteratively identified by the query strategies. To evaluate each query strategy, the initial training set was used to train a classifier, and then the particular query strategy was used to select the most informative points for labelling. While in an actual experiment, labelling of a selected point by the Oracle would require weeks of wetlab experiments, in our simulation labelling simply consisted of looking up the withheld previously known class of an instance. The chosen (i.e. newly labelled) instances were subsequently added to the training set, and removed from the unlabelled pool of asparagine and aspartate. The model was retrained with the new training set. The static test set was used as a holdout test to estimate the classification performance over time, as instances were labelled and added to the training set. A better active learning strategy will result in improved classification performance while requiring fewer labelled training data instances. This will be measured via performance curves as detailed in section 3.4.

In this chapter, we present the protein datasets collection, feature selection, generation of initial training data set, model selection, and the proposed pool-based active learning framework with SVM as it relates to binary classification problems. We present performance evaluation of the various active learning query strategies considered through learning curves. Finally, we explore how to determine when to stop active learning challenge.

3.2 Datasets Collection, Model Selection and Generation of Initial Training Set

The dataset used for the evaluation of active learning algorithms was taken from experimentally verified and validated N/D hydroxylation sites of 41 proteins known to have at least one N/D hydroxylation site. (Liu, 2009) extracted this dataset from the human protein entries in dbPTM, Swiss-Prot, PhosphoELM, and O-GLYCBASE. Following (Liu, 2009), feature vectors consisted of surface accessibility (SA), secondary structure (SS), and position-specific scoring matrix (PSSM) data. Here the surface accessibility features were generated by RVP-NET; a neural network computational model that predicts solvent accessibility value from sequence information of each protein target. Then, the secondary structure features were determined by parallel cascade identification (PCI-based) protein secondary structure prediction, which distinguishes between ά-helices, β-strands and non-regular structural elements from primary sequence data (Liu, 2009; Green et al., 2009). The PSSM features were obtained using PSI-BLAST, which calculates position-specific scores from a multiple sequence alignment of the of target protein with sequence-similar proteins (Altschul et al, 1997). These scores capture sequence conservation at each amino acid position of the protein. All feature data was converted to numerical values for input to the SVM classifiers. The details regarding the computation and compilation of these feature vectors can be found in (Liu, 2009).

Taken together, these features capture both the physiochemical and biochemical properties for SA and SS as well as sequence conservation information for the PSSM for the 15 amino acids surrounding each asparagine/aspartate site (i.e., window of ±7 amino acids (AAs), centred on the N/D was used). This resulted in feature vectors having 360 dimensions (surface accessibility = 15, secondary structure = 45, and PSSM = 300). It should be noted that the feature vectors are expected to be somewhat inaccurate since all input feature data (i.e. SS, SA, PSSM) are calculated by other tools which may have inherent errors. All identical feature vectors were removed from the dataset and the remaining 1813 non-identical data samples, with 55 positive and 1758 negative data points (numeric values), were used to induce SVM-based active learning simulation.

The libSVM toolbox for MATLAB was used with the radial basis function (RBF) kernel function $K(x,y) = \exp(-\gamma||x-y||^2)$ in all active learning experiments as other functions considered did not show improved performance in comparison to RBF.

We performed model selection to tune the libSVM classification parameters (i.e., the slack penalty coefficient C, and the RBF kernel parameter (γ)) in order to control the trade-off between classifier complexity and the misclassification rate over the training samples (Gold & Sollich, 2002; Chapelle & Vapnik, 2000). The optimal parameters (C = ***32768***, γ = ***0.0156)*** were obtained by performing a parameter sweep over the range $C = (2^{-5},\ldots, 2^{15})$ and $\gamma = (2^{-15},\ldots, 2^{3})$ using 10-fold cross-validation over all 1813 data points, as done in (Hsu et al., 2010). These parameters led to a cross-validation accuracy rate at 98%.

For the dataset split, 100 (20 positive and 80 negative) instances were randomly drawn to create the test set for uncertainty-based, certainty-based, density-based and passive query strategies. An initial training set of 3 positive and 3 negative instances was drawn from the remaining data (i.e. 35 positive and 1678 negative instances). For uncertainty-based, certainty-based and passive query strategies, the initial training set was drawn randomly, whereas for density-uncertainty-based query strategy, the 3 most representative instances each from the 35 positive and 1678 negative classes were selected by K-means clustering to form the initial training set. The remaining data (i.e. 32 positive and 1675 negative instances) formed the set of unlabelled instances used to induce the active learning process. The test data set used for the performance evaluation of the active learning was the same across all experiments for each of the query strategies considered. Therefore, the static test set was never used in the development process of the active learning cycle. Although, it would be desirable to have a truly independent test set to be used solely for evaluation of the final solution, the amount of data available was limited by the positive class. Such a true independent test set would be required to obtain an unbiased estimate of the true error rate of the final classifier.

3.3 A Pool-Based Active Learning with SVM in Binary Classification

In the active learning procedure, the active learner interactively queries instances for learning the predictive model. For a binary classification problem, we begin with a small initial training pool of labelled instances L= $\{x_1, \ldots, x_k\}$ and typically much larger pool $U = \{x_{k+1}, \ldots, x_k\}$ of unlabelled instances, where each instance x_i is a vector in some feature space X.

Each instance x_i has an associated true class or label $y_i \in \{-1, 1\}$, which is initially known to the classifier for points in L. In the pool-based active learning sampling, the *active learner*, ℓ, consists of two components: (C, Q), where C denotes the classifier (SVM) learning algorithm and Q is the query function *(i.e., the selection function)* that determines the most informative unlabelled instance in U (Lewis & Gale, 1994; Roy & McCallum, 2001) as previously indicated in Figure 2.6 (c) to be labelled by the Oracle. Active learning is an iterative process as shown in Figure 3.1.

Procedure: Active Learning with query strategy $(L, U, instance, x)$

Input: Let L be initial small training set; U the pool of unlabelled data set

Output: Labelled training set L, final classifier C

1. Use L to train the initial classifier C (SVM)
2. **Loop** while adding new instances to L
   - Use the current classifier C to predict label of all unlabelled instances in U
   - Use query strategies Q to select most informative unlabelled instances x_t from unlabelled pool U, and query Oracle for true label y_t
   - Add newly labelled x_t to L, and remove from U
   - Use L to retrain the current classifier C
   - Evaluate the classifier C's performance on an independent test set after each query
3. **Until** the predefined stopping criteria is reached or all unlabelled data has been
4. **Return**
5. **Report** the average of the results of all tests

Figure 3.1. Proposed Pool-Based Active Learning with Query Strategy.

In each learning cycle, the classifier was trained on all available labelled training data, and then applied to the unlabelled data. The current classifier then applies Q (i.e., the selection or query function) to choose one unlabelled instance x_t from U considered to be most informative (Xu et al., 2009; Zhu et al., 2010) for annotation by the Oracle. The pair (x_t, y_t) is added to L and removed from U. The classifier is then retrained and the process is repeated until a predefined stopping criterion is reached or all unlabelled instances have been selected and labelled (Baram et al., 2004; Liu, 2004).

Tong & Koller (2000) noted that the effectiveness of active learning is dependent on the query strategy's ability to acquire the most useful instances at each iteration. We utilized the pool-based active learning uncertainty query strategy (Lewis & Gale, 1994), density-uncertainty query (Mohamed et al., 2010; Xu et al., 2007), and certainty-based (Liu 2004; Warmuth et al., 2002) query strategies with libSVM for a binary classification problem. Then unlabelled instance closest to the separating hyperplane was selected by uncertainty and density-uncertainty query strategies according to the libSVM formulation, as described below.

3.3.1 Uncertainty Measure of Unlabelled Instances

To measure the uncertainty/certainty of an unlabelled instance for each query strategy, we use the probability estimate available from the libSVM toolbox (Chang & Lin, 2011). Here, a probability of 1.0 indicates the maximum probability of being a positive instance, while a

probability of 0.0 indicates the maximum probability of being a negative instance. An unlabelled instance whose probability estimates falls closest to 0.5 is considered as the most informative instance (i.e., most uncertain) for the uncertainty-based and density-uncertainty-based query strategies. The probability estimate is closest 1.0 is selected by the certainty-based query strategy for annotation. An intuitive illustration of this representation is shown in Figure 3.2.

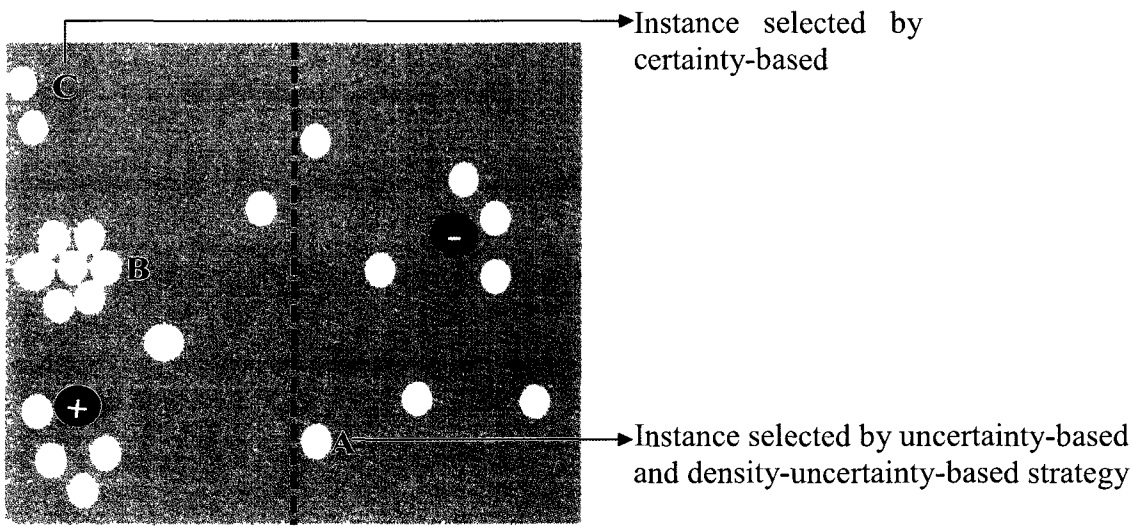

Figure 3.2. Representation of informativeness measure (adapted from Tomanek (2010)).

The most uncertain instance (A) is an ideal candidate instance for the uncertainty-based and density-uncertainty-based query strategies compared to the highly representative, but less informative, instance (B). For certainty-based, the most informative instance will be (C) (i.e., far from decision boundary on the positive side). The dashed line represents the current decision boundary. The small white circles are unlabelled instances, while the purple and green circles are labelled training instances.

3.4 Proposed Active Learning Pool-Based Framework

The Figure 3.3 emphasizes the input space visualization of the active learning process. The flowchart in Figure 3.4 illustrates the idea of selective sampling process in SVM pool-based active learning algorithm for the current study. Each experiment begins with splitting the data, i.e., the set of all labelled N/D sites, into the initial training set, a static test set, and the pool of unlabelled instances. We build a model using the initial labelled training set. Then, using the appropriate query function strategy, we select an unlabelled instance from the pool and asked the Oracle for the true annotation of the queried unlabelled instance. The chosen instance is subsequently added to the training set, then, the model is retrained with the new training set. At each iteration, the model's classification performance is evaluated using the static test set. The process is repeated until a predefined stopping criterion is reached or all unlabelled instances are exhausted.

3.4.1 Learning Curves

The relationship between the number of labelled instances and a model's performance can be visualized by learning curves that show the classification performance as a function of annotation effort. At each iteration of active learning, the model's performance is evaluated using a hold-out static test set. More complex resampling-based evaluation techniques such as cross-validation, leave-one-out, or bootstrap cannot be used to estimate the performance of an active learner directly from the pool of labelled instances chosen by the learner. The reason being that instances selected for labelling by a good active learner tend to be heavily biased towards 'hard' instances that do not reflect the true underlying distribution of the entire dataset (Tomanek 2010; Kang et al., 2004).

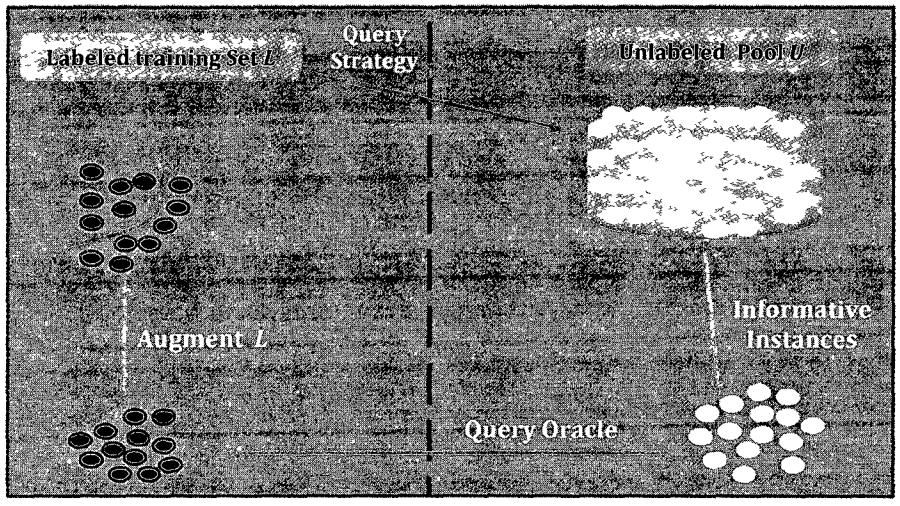

Figure 3.3. Active Learning Input Space Visualization (reproduced from Singh, 2008).

The framework in Figure 3.3 uses the following steps to accomplish the selective sampling process. Here, the classifier is trained with small labelled training set, *L*. Then, the query strategy selects the most informative instances and asks the Oracle for its label. Finally, the training data is augmented with newly labelled instance.

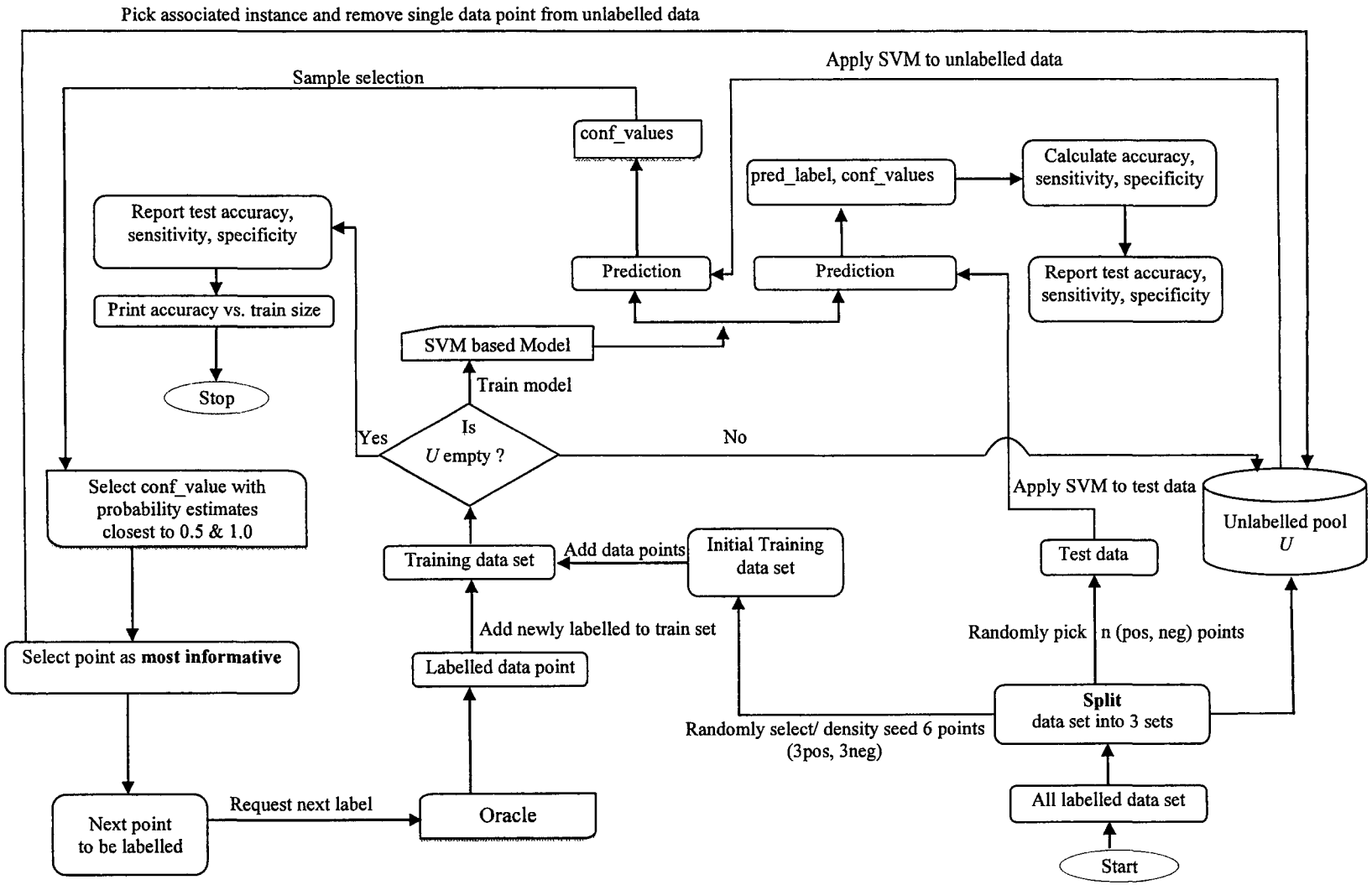

Figure 3.4. Framework of sample selection for a pool-based active learning.

3.4.2 Performance Evaluation Measures (Metrics)

Performance evaluation metrics play an important role in the assessment of active learning performance. A variety of common metrics are defined based on the *confusion matrix* (also called contingency table), two-by-two confusion matrix shown in Figure 3.5.

		True class	
		Positive	Negative
Predicted class	Positive	True Positives (TP)	False Positives (FP)
	Negative	False Negatives (FN)	True Negatives (TN)

Figure 3.5. Confusion matrix for a binary classification.

TP (True Positives) are the number of correctly predicted positive instances, TN (True Negatives) are the number of correctly predicted negative instances, FP (False Positives) are the number of instances incorrectly predicted to be positives and FN (False Negatives) are the number of instances incorrectly predicted to be negative.

In this thesis, the performance was measured using sensitivity (recall), positive predictive value (PPV), Matthews' correlation coefficient (CC), and area under the receiver operating characteristic curve (AUC) as defined below. Other relevant evaluation metrics used to check the distribution of selected instances are the number of selected positives, the number of selected negatives, the number of selected decision values, the number of support vectors and the distance of a selected point to the hyperplane.

Sensitivity (recall) is the percentage of true positive prediction that is correctly detected by the classifier.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Precision of a classifier is the percentage of positive predictions made by the classifier that are correctly predicted. It is also called predictive positive value (PPV).

$$\text{Precision (PPV)} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Matthews Correlation Coefficient (CC): The CC is used as a measure of quality for binary classifications and is relatively unaffected by class imbalance in the test set. The CC in essence is a correlation coefficient between the observed and predicted binary classification results; it returns a value between −1 and +1, where a coefficient of +1 represents a perfect prediction, 0 an average random prediction, and −1 an inverse prediction.

$$\text{CC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

AUC: Area under the curve is a measure of classifier's discriminatory performance that shows how successfully and correctly a classifier separates the positive and negative observations in binary classification settings. Since the AUC metric evaluates the classifier across the entire range of decision thresholds, it gives a good overview of performance of the model across classes' distribution (Seyda et al., 2007). We have used AUC to summarize the performance of a classifier into a single metric for each active learning query strategies and passive learning

technique considered. AUC values lie in the range [0, 1]; the closer AUC is to 1, the better the overall classification performance on an independent test set.

3.4.3 Results and Discussion

To verify the effectiveness of the active learning, we evaluated and compared the performance of the various active learning techniques with passive learning as a function of a growing labelled training set size, and visualize the performance as learning curves. All experiments were repeated 20 times (i.e. 20 runs) to account for the stochastic nature of the algorithms.

Tables 3.1 and 3.2 show the results obtained by each strategy after 50 and 200 instances were added to the training set respectively. For each, the mean and standard deviation observed over the 20 runs is reported. Tables 3.1 and 3.2 clearly demonstrates the consistently better performance achieved by the density-uncertainty-based strategy in terms of recall, PPV, CC and AUC. In fact, all active learning strategies outperform passive learning on all measures except for PPV for the certainty-based strategy due to its bias towards positive predictions as discussed below.

The results in Table 3.1 are similar or even better than that in Table 3.2. This indicates that adding 200 instances to the training set do not bring any significant improvements over 50 instances under the same experimental settings. These observations showed that active learner trained on fewer instances could attained the same or even higher performance than one trained on the whole dataset (Schohn and Cohn 2000; Ertekin et al., 2007; Vlachos, 2008).

Table 3.1. Average performance results of different active learning query strategies after 50 queries measured over 20 runs (± standard deviation of the average).

Query Strategies	Average measures + standard deviation					
	Recall	**PPV**	**CC**	**AUC**	**Num_Pos**	**Num_Neg**
Uncertainty-based	96% ± 0.03	98% ± 0.04	0.87 ± 0.03	0.98 ± 0.00	22.4 ± 0.28	32.5 ± 0.30
Certainty-based	100% ± 0.00	80% ± 0.02	0.83 ± 0.20	0.98 ± 0.00	35.0 ± 0.00	20.0 ± 0.14
Density-uncertainty based	97% ± 0.02	100% ± 0.00	0.90 ± 0.02	0.98 ± 0.00	21.5 ± 0.28	33.6 ± 0.28
Passive-based	90% ± 0.03	97% ± 0.05	0.81 ± 0.04	0.95 ± 0.03	4.1 ± 0.14	50.6 ± 0.14

Table 3.2. Average performance results of different active learning query strategies after 200 queries measured over 20 runs (± standard deviation of the average).

Query Strategies	Average measures + standard deviation					
	Recall	**PPV**	**CC**	**AUC**	**Num_Pos**	**Num_Neg**
Uncertainty-based	98% ± 0.00	100% ± 0.00	0.89 ± 0.01	0.98 ± 0.00	34.0 ± 0.14	171.0 ± 0.07
Certainty-based	100% ± 0.00	88% ± 0.01	0.87 ± 0.10	0.97 ± 0.01	34.0 ± 0.35	170.9 ± 0.35
Density-uncertainty-based	96% ± 0.01	100% ± 0.00	0.90 ± 0.01	0.99 ± 0.00	35.0 ± 0.07	168.8 ± 0.07
Passive-based	84% ± 0.02	100% ± 0.03	0.77 ± 0.02	0.96 ± 0.01	6.3 ± 0.14	198.6 ± 0.14

This behaviour may be due to the limited amount of positive data available, as there are only 55 positive points overall, 3 of them are selected for the initial training set, and an additional 20 are selected for the static test set leaving only 32 positive points in the unlabelled set L. By the 50^{th} iteration of active learning, all the positive points have likely already been added to the training data. This is particularly true for the certainty-based approach which favours positive instances. Therefore, between 50 and 200 only negative points are likely to be added to the labelled training set. The performance evaluation learning curves for the following metrics will only be plotted up to 50 iterations (queries): recall, PPV, CC, AUC, number of positives and negatives. Later in the chapter, 200 iterations will be shown for the distribution of unlabelled instances, number of support vectors, and distance of an instance to hyperplane within SVM margin.

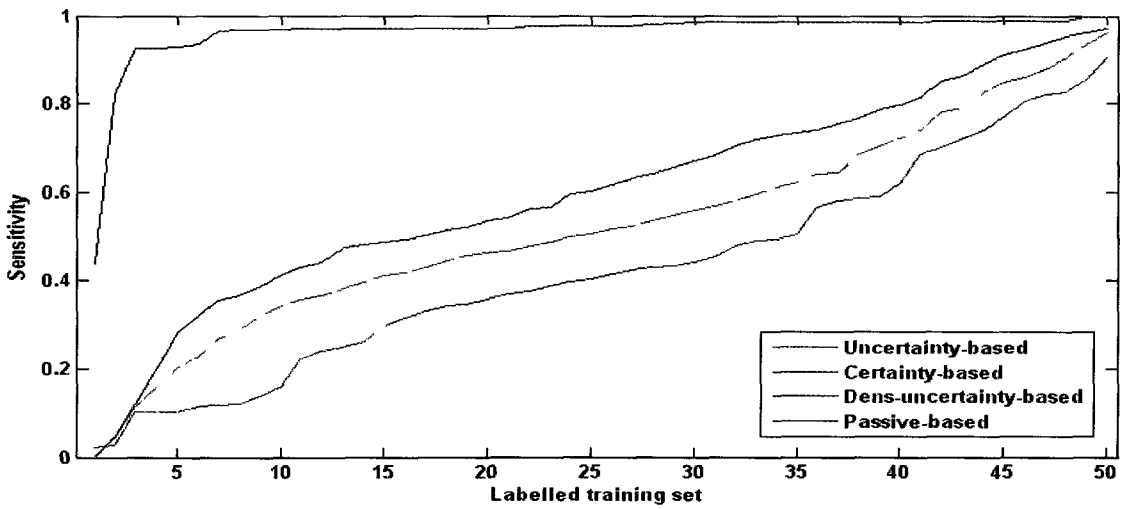

Figure 3.6. Learning curves for evaluation of recall.

The X-axis is the number of labelled instances in the training set and the Y-axis represents the recall values for active and passive learning strategies.

Figure 3.6 shows the performance results in terms of sensitivity (or recall) for the three active learning strategies and the baseline passive learning strategy as the size of the training set increases at each learning cycle (i.e., iteration). As expected, the certainty-based strategy quickly attains 98% ± 0.00 recall at only 30 labelled instances since it adds primarily positive instances to the training data, biasing the entire classifier towards positive prediction. The density-uncertainty-based and uncertainty-based query strategies were remarkably good and achieved recall values of 97% ± 0.02 and 96% ± 0.03 after the 50 labelled instances. The density-based query was slightly higher than uncertainty-based due to a better initial training set as chosen by K-means clustering . The passive learning approach achieves relatively poor recall and grows slowly with training set size. This is due to the predominance of negative data points in the unlabelled set due to the overall class imbalance observed in the total dataset. Therefore, the passive approach is most likely to select negative instances at each iteration, biasing the classifier towards negative predictions. The active learning strategies clearly address this class balance, leading to excellent classification performance (Ertekin et al., 2007). In terms of recall, all active learning query strategies outperformed passive learning.

Figure 3.7 shows PPV for all algorithms. The PPV rates for density-uncertainty and uncertainty-based strategies follow a similar pattern of rapid growth in the first 10 instances followed by relatively slow growth to 100% ± 0.00 and 98% ± 0.04 completions respectively. The strong PPV is due to the effectiveness of the active learning query strategies at correctly identifying the most informative hydroxylated sites for annotation in each iteration. The certainty-based strategy shows rapid growth in the first 4-5 labelled instances, but then grows more slowly than both uncertainty-based strategies. This mirrors the recall curve show in Figure 3.5 above, where the

classifier performance over positive test instances initially improves with more positive instances added to the training set, but quickly saturates. The trade-off between recall and PPV is clearly observable with certainty-based having lesser PPV value but a higher recall than the passive learning strategy. In terms of PPV, both density-uncertainty and uncertainty-based strategies were consistently superior to the passive learning (random selection).

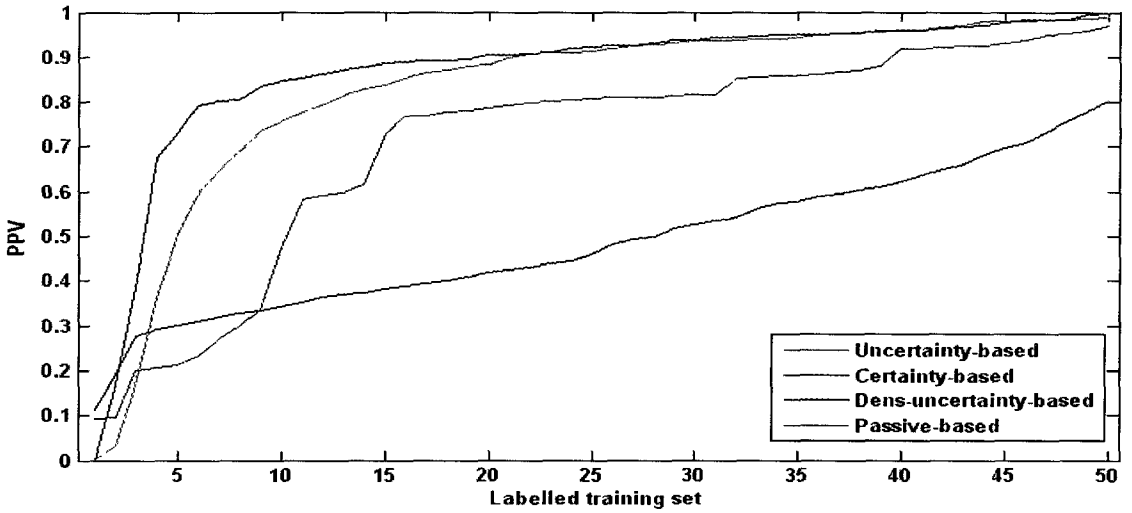

Figure 3.7. Learning curves for evaluation of PPV.

The X-axis is the number of labelled training set and the Y-axis represents the PPV values for active and passive learning.

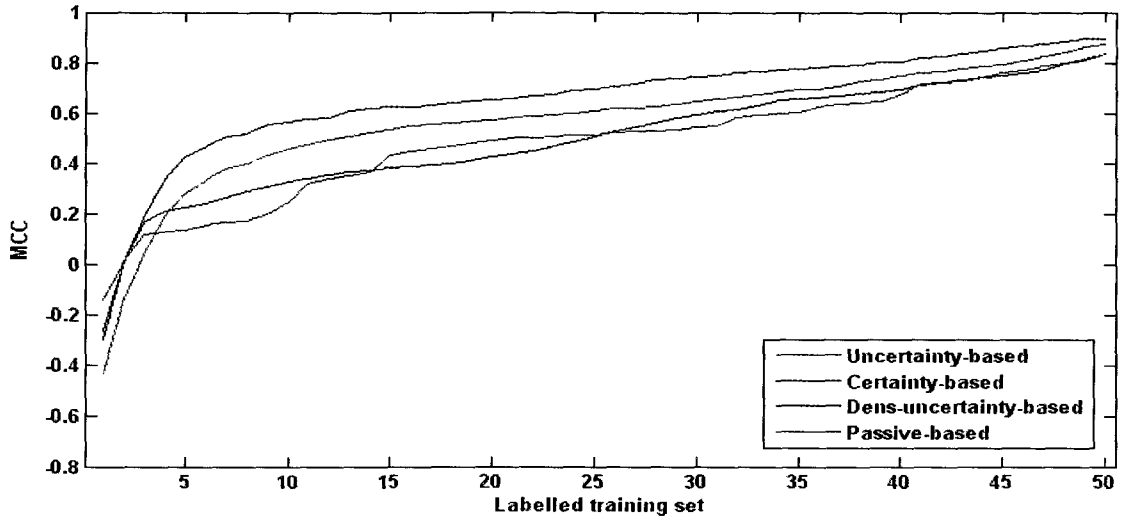

Figure 3.8. Learning curves for Matthews' CC evaluation.

The X-axis is the number of labelled instances in the training set and the Y-axis represents the CC values for active and passive learning.

Figure 3.8 shows the performance correlation coefficient values for the three active learning strategies and the passive learning strategy. All active learning strategies outperform passive learning in terms of Matthews' CC, which combines both sensitivity and PPV into a single measure. Although the certainty-based strategy appeared to be superior when performance was measured solely in terms of recall, it is now clear that this is, in fact, a sub-optimal strategy when false positives are also considered. Both uncertainty-based strategies perform very well here, however the initial advantage of the density-uncertainty strategy is clearly evident. For example, to reach a CC of 0.5 requires only 7 labelled instances for density-uncertainty and 13 labelled instances for uncertainty-based, however this performance level is only obtained after 22 instances for the passive query strategy.

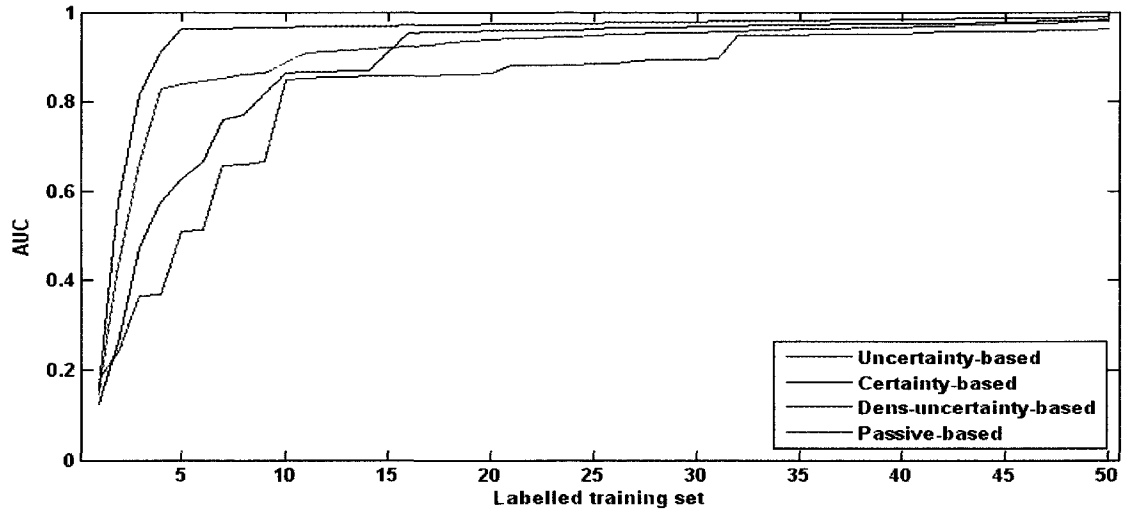

Figure 3.9. Learning curves for AUC evaluation.

The X-axis is the number of labelled instances in the training set and the Y-axis represents the AUC values for active and passive learning.

The learning curves in Figure 3.9 show the performance evaluation results of the various strategies in terms of AUC. Like PPV and Matthews' CC, AUC is a combination of both sensitivity and specificity. Therefore, it demonstrates much the same trends, with the density-uncertainty-based approach showing an early lead over all methods rising quickly to attain AUC=0.96±0.00 with only 4 queries, however beyond this point there was no significant improvement. Uncertainty-based query strategy reaches 0.82 ± 0.02 AUC values at the same 4 queries, while certainty-based query attained only 0.57 ± 0.01 AUC at the same iteration. In contrast with the active learning strategies, at 4 labelled instances, the passive learning strategy had just achieved only 0.36 ± 0.03 AUC value, but it does achieve the remarkable AUC value of 0.96 at the end of the learning cycle. This also demonstrated again that all active learning performs better than passive learning. It is interesting to note that the uncertainty-based strategy

appears to outperform the certainty-based strategy in the first 5 iterations. This would indicate that the classifier is able to differentiate between positive and negative sites, however the decision threshold is admitting too many false positive predictions (i.e. a more stringent threshold may lead to lower false positive rates while only sacrificing a small degree of recall).

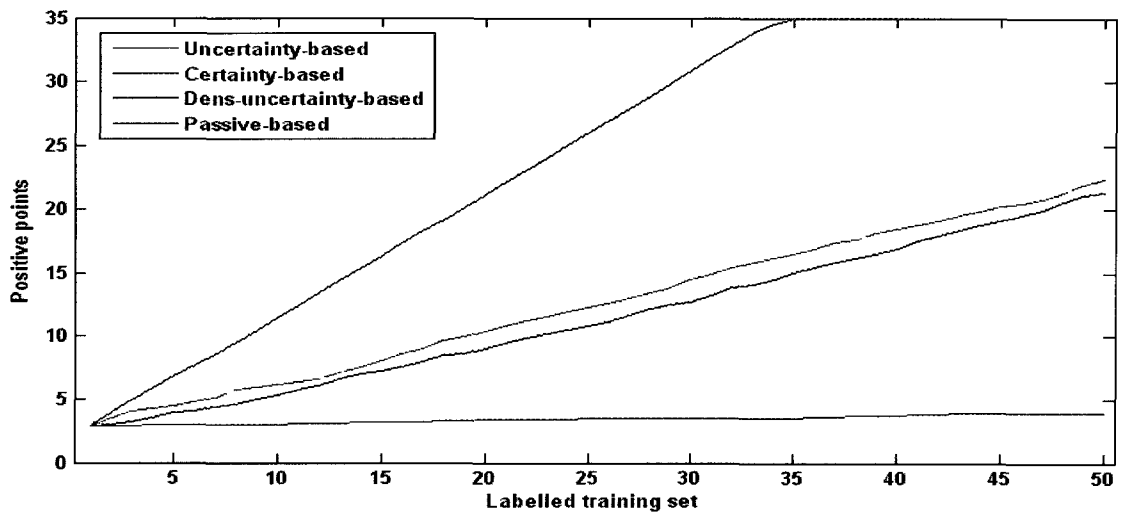

Figure 3.10. Learning curves for number of selected positive instances.

The X-axis is the total number of labelled instances in the training set and the Y-axis represents the number of positive instances selected for labelling for active and passive learning.

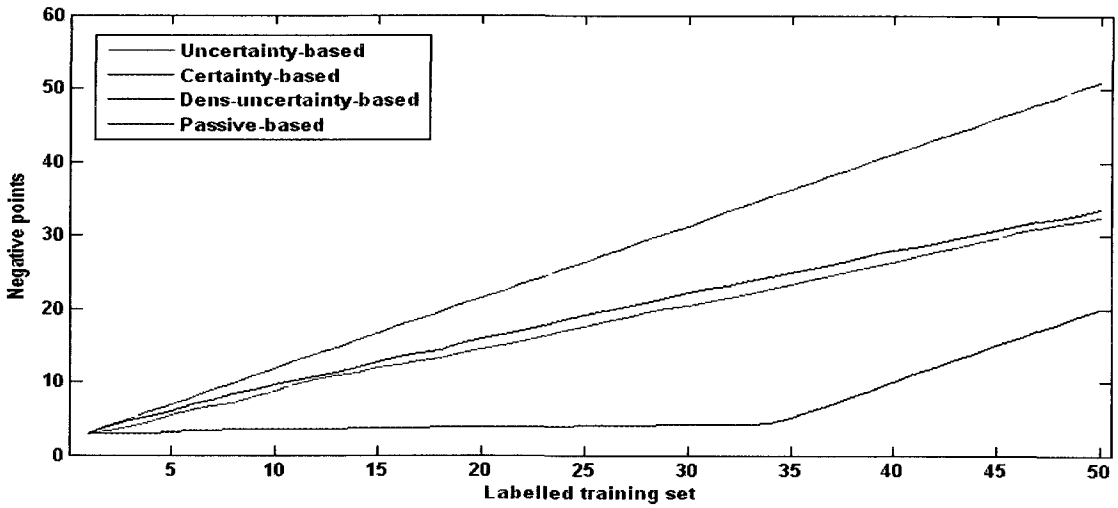

Figure 3.11. Learning curves for number of selected negative instances.

The X-axis is the total number of labelled instances in the training set and the Y-axis represents the number of selected negative instances for active and passive learning.

Figure 3.10 shows the number of positives selected by each strategy at every round of the learning process, while Figure 3.11 illustrates the number of negative instances selected for annotation. As expected, the certainty-based sampling chooses purely positives at the beginning, yielding a 45° slope until all positives in the unlabelled instances pool, *U*, were exhausted. This is similar to the behaviour observed by Forman (2002) with active learning to incrementally select only the strongest positive prediction from unlabelled instances with SVM.

Subsequently, the classifier begins selecting negative instances for labelling, and the curve quickly flattens out. In both uncertainty-based and density-uncertainty-based active learning query functions, the number of positives grew linearly, but with a lesser slope than certainty-

based learning. This indicates a more balanced selection of points as sometimes the most informative point is a positive instance, and sometimes it is a negative instance. In the passive learning technique however, the learning curve differs significantly. Since instances were randomly selected from an unlabelled pool that was, in fact, overwhelmingly biased towards negative instances (1675 negative points vs. 32 positive points), the passive learner naturally selected negative instances almost exclusively. With no guidance on which instance was most informative, the passive learning algorithm is unable to deal with the large class imbalance for this application, selecting only 4.1 ± 0.14 positive instances at the end of the learning process. In contrast, after 50 iterations, certainty-based sampling identified all 35.0 ± 0.00 positives, uncertainty-based sampling identified 22.4 ± 0.28 positives, and density-uncertainty based sampling identified 21.3 ± 0.28 positives from the unlabelled data points.

When all the performance metrics are taken together, the results strongly demonstrate the usefulness of active learning over passive learning. In other words, active learning query strategies outperformed passive learning primarily due to careful selection of instances for labelling. The ability of active learning to identify positive training instances among the pool of unlabelled instances in the face of a severe class imbalance is also a strength of active learning. This is explored in greater detail in the next section.

3.5 SVM Performance Evaluation within the Margin

In the SVM-based classification problem, the ultimate goal is to find an optimal separating hyperplane that maximizes the distance between the hyperplane and the closest training instances

(i.e. the decision margin); thereby minimizing the number of errors over the training instances and promoting generalization to new future test data (Rosset et al., 2003).

In this section, we examine the ability to overcome class imbalance in the unlabelled dataset by selecting instances for labelling from within the SVM margin surrounding the decision hyperplane. The uncertainty-based and density-uncertainty-based active learning approaches both select instances for labelling that fall close to the decision hyperplane, and therefore, these methods are likely to sample instances from within the margin. Furthermore, we also looked at the influence of the most informative instances on support vectors and estimation of the distance of a point to the hyperplane in order to reemphasize the significance of margin in active learning. Experiments in previous works have shown that margin-maximization usually leads to a reasonable prediction model (Breiman, 1999; Grove & Schurumans, 1998).

Our experimental results above have clearly demonstrated that the uncertainty-based active learning techniques achieve a better performance over the passive learning even with the class distribution imbalance of the unlabelled dataset. These performance results reaffirmed the usefulness of the selection strategy that focuses on the most informative unabelled instances within the margin. Since, in an SVM, the hyperplane is placed equidistant from the positive and negative classes, the unlabelled instances that fall within the margin will tend to have a more balanced class distribution than that of the entire dataset. Hence, instance selection within the margin can address the data class label imbalance issues during the active learning process.

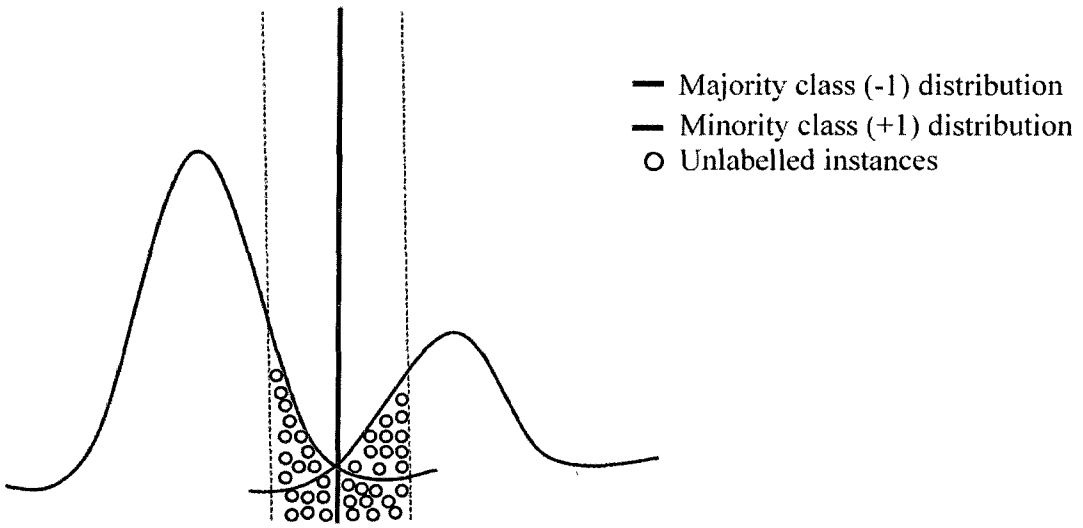

Figure 3.12. Instances within the margin are less imbalanced than the entire dataset (reproduced from (Ertekin et al., 2007)). The lower region of the margin-hyperplane diagram is filled with unlabelled instances within the margin. Dashed lines indicate the margin, and the solid is the separating hyperplane.

Suppose that the class distribution of an imbalanced dataset is given in Figure 3.12 (Ertekin et al., 2007). The class imbalance ratio within the margins is much lesser than the entire dataset imbalance ratio. As mentioned above, in the presence of a strong class imbalance (in this case towards the negative class), the passive approach will tend towards selecting only instances for labelling that come from the dominant class. Then, using active learning to select the most informative instances within the margin accounts for the better performance by active learning than passive learning where instances to be labelled were randomly selected from the entire data pool.

3.5.1 Support Vectors Evaluation

SVM usually has large number of support vectors, especially when the feature value distributions of the positive and negative training instances are highly overlapped (Zhan & Shen, 2005). However, if the classes do not overlap too much in the feature space, the number of support vectors will be small with respect to the training set size (Osuna & Girosi, 1998).

In Figure 3.13, we investigated how the number of support vectors changes with each iteration of active or passive learning in our experiments. The learning curves for uncertainty-based and density-uncertainty-based active learning algorithms are drastically different from certainty-based and random-based sampling techniques. This is because; both uncertainty-based and density-uncertainty-based query functions intelligently picked unlabelled instances closest to the hyperplane within the margin of the SVM-based classifier. The annotation of these instances enriched the SVM-based classifier by providing new support vectors required to maximize the margin. Support vectors within the margin directly influence the shape of the hyperplane (Tong & Koller, 2001). Instances which are correctly predicted but lie outside the decision margin (i.e. instances which are not likely to become support vectors) do not contribute to constructing the optimal separating hyperplane; even if their position is changed, the hyperplane and margin will remain the same.

As correctly depicted in Figure 3.13, there is an increase in number of support vectors used by both active learning query strategies with a corresponding increase in the number of training set. This learning curve illustrates that a greater number of newly labelled instances are being leveraged to optimize the decision boundary (i.e. more newly labelled instances are found to be

support vectors) for the uncertainty-based strategies when compared with passive and certainty-based strategies.

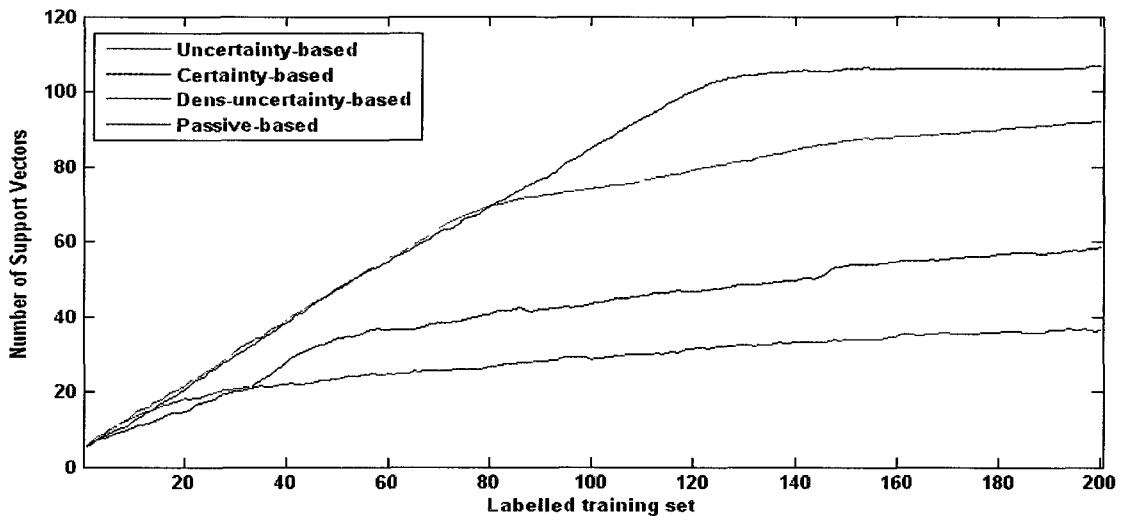

Figure 3.13.Learning curves for selected support vectors.

The X-axis is the number of labelled instances in the training set and the Y-axis represents the number of support vectors used by the trained model at that iteration.

Support vectors learning curves for uncertainty-based and density-uncertainty-based query functions are roughly with unit slope until approximately 70 and 110 support vectors have been selected. Beyond this point, very few support vectors are added. However, the certainty-based sampling does not exhibit such behaviour because it is only interested in unlabelled instances farther from the hyperplane. Being further from the hyperplane, these newly labelled instances are far less likely to become support vectors and influence the shape of the hyperplane. Since only instances which are support vectors are actually used to define the decision boundary hyperplane, expending resources labelling instances (via wetlab experiments) is wasteful if they

are not going to influence or modify the definition of the decision boundary hyperplane. Likewise, the passive learning strategy adds support vectors at a similar rate since instances are chosen at random for labelling, rather than focussing on those points that are likely to become support vectors.

3.5.2 Distance of a point to hyperplane evaluation

Here, we discuss the relationship between the average distance of a point to the current hyperplane and increasing number of labelled instances. Further, we examine how the selection of unlabelled instances closest to the hyperplane leads to improved performance. As the learning continues the margin becomes larger, which ultimately leads to a better generalization ability of the classifier and prevents overfitting (Bennett & Demiriz, 1999). This implies that the most informative unlabelled instances are found within the margin. LibSVM software was retuned in consultation with the LibSVM developers (Chang & Lin, 2011) to be able to calculate the distance of a point to the hyperplane, as this was not previously provided by the original libSVM outputs.

However, as the learning cycle continues with a corresponding increase in the training set, it becomes difficult for the classifier to find instances within the margin most likely because the unlabelled instances closest to the hyperplane are now outside the margin band. This could apparently leads to an empty margin band, which indeed appears to be a potential stopping criterion for active learning in order to avoid labelling less informative unlabelled instances or outliers (Campbell et al., 2000). Intuitively, we could assume that an SVM-based classifier could

be transformed from soft-margin into hard-margin, where no support vectors are allowed within separating margin as the SVM classifier maximized the separating margin.

Figure 3.14 shows a direct relationship between the average distance of the unlabelled instance selected for labelling to the current hyperplane versus the number of labelled instances. With the density-uncertainty query strategy, the learning curve initially increases with a corresponding increase in the number of the training sets. This query strategy quickly selects all the instances that fall within the final classifier margin and then the average distance suddenly drops, followed by a more gradual decrease as the learning continues. This is referred to as the "*rise-peak-drop* flattening pattern (Vlachos, 2008). It rises at the beginning; then reaches its maximum after which it gradually drops and finally flattens out for the reasons discussed above.

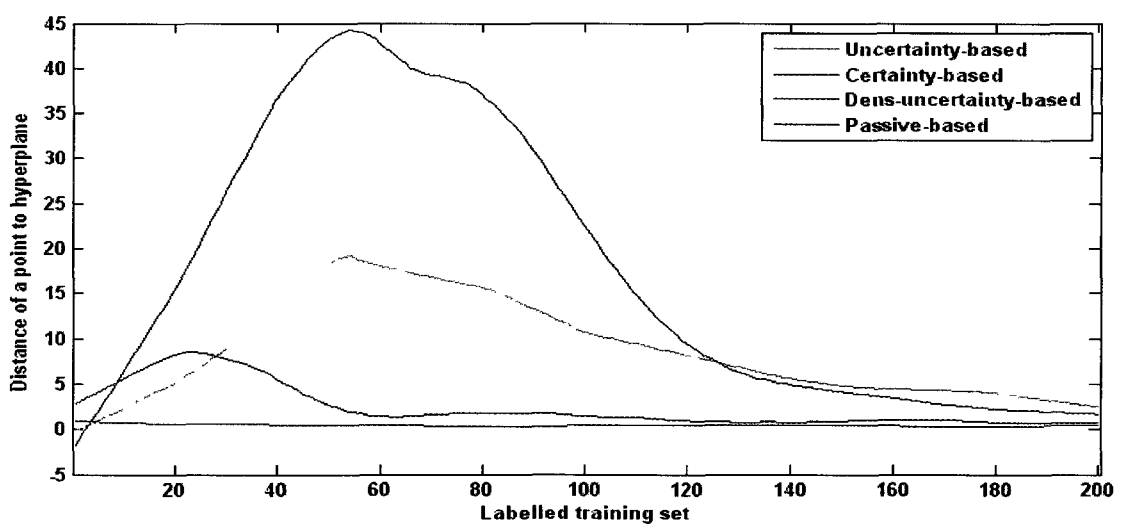

Figure 3.14. Learning curves for selected unlabelled instances distance from the hyperplane.

The X-axis is the number of labelled instances in the training set and the Y-axis represents the average distance of the instance selected for labelling to the separating hyperplane for active and passive learning.

This is because the distance of the SVM hyperplane from the nearest instance on either side is maximized as the most informative instances within the SVM margin are being labelled. That is, those instances closest to the margin are quickly selected for labelling, and are removed from the unlabelled set, thereby increasing the distance to the nearest unlabelled instance. The initial improvement could be attributed to the increasing distance of the instances from the hyperplane as the SVM classifier maximizes the margin (Ratsch & Warmuth, 2003) as discussed above. The learning curve indicates the closest unlabelled instances to the hyperplane are outside the margin band after 50 labelled instances (queries) for both uncertainty-based query strategies and certainty-based query because of a similar *"rise-peak-drop"* flattening pattern, but to a lesser extent. Passive query on the other hand, does not seem to portray such behaviour as instances are randomly selected for labelling.

3.5.3 Selected unlabelled decision values within the margin

We further explored the behaviour of the decision value of those unlabelled instances selected for labelling during each round of the learning cycle to further examine the behaviour of active learning with respect to the SVM margin. Decision values are output from the SVM for each instance, where a value of 1.0 indicates that the instance is most likely to be positive ideal for certainty-based query, while a decision value of 0 indicates that it is most likely negative. Then,

a value of 0.5 indicates that the classifier is completely uncertain of the class membership for this instance, most useful for uncertainty-based queries.

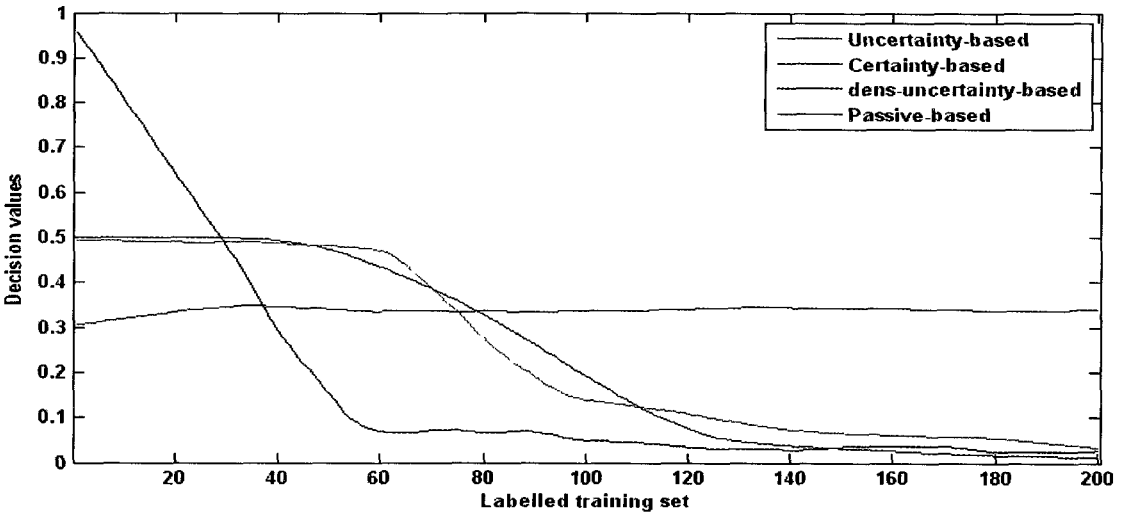

Figure 3.15. Learning curves illustrating the SVM decision values. X-axis is the size of the labelled training set. The Y-axis represents the selected unlabelled decision values during each learning step (iteration) for active and passive learning.

Figure 3.15 shows a direct relationship between selected unlabelled instance decision values and growing training set size. It is observed in Figure 3.15 that both density-uncertainty and uncertainty query strategies quickly identify all the most informative instances (i.e. those instances with decision values close to 0.5) within the margin. After 50 labelled instances or queries, both strategies began to select less informative instances and consequently led to a gradual drop in the curves and a flattening out. There is strong connection between the selected unlabelled instance decision values and the average distance of the selected unlabelled instances

closest to the hyperplane for uncertainty-based queries in Figure 3.13. The certainty-based strategy exhibited a steep drop as it selects all points believed to be positive in descending order of certainty (i.e. from 1.0 down).

The behaviours shown by all three query strategies indicated that these classifiers became highly saturated at some point, and adding more instances will not contribute to finding a unique optimal decision boundary. On the other hand, passive sampling only selected instances at random. Therefore, the decision values of these randomly selected instances are not expected to show a trend over time. This consistent behaviour shown by both query strategies could be used to deduce a stopping criterion for active learning, as the idea is similar in concept with the stopping criterion discussed in Schohn and Cohn's (2000) of stopping annotation when there are no instances within the SVM margin closest to the separating hyperplane.

3.6 Summary

In this chapter, we presented results of the active learning experiments based on the proposed active learning framework and the baseline passive learning strategy. The performance results achieved by the various query strategies and passive learning were expressed as learning curves. We have also observed how training with fewer but more carefully selected instances achieved a similar or even higher performance than those trained on a larger asparagine/aspartate dataset as shown in Table 3.1 and 3.2. The experimental results showed how density-uncertainty-based strategy has consistently outperformed other active and passive learning strategies. We also presented and demonstrated that selecting unlabelled instance within the classifier's margin leads to improved classifier performance. Therefore, a stopping criterion could be deduced from such

behaviour exhibited by uncertainty-based active learner, as it explicitly became evident that viable unlabelled instances ceased to exist once all the most informative or useful instances within the margin have been selected for labelling.

3.7 Conclusion

We have evaluated various active learning query strategies, which intelligently select the most informative unlabelled instances for annotation by the Oracle, and compared them to a baseline passive learning strategy. We proposed and integrated these active learning query strategies with an SVM classifier for the prediction of N/D hydroxylation site on proteins. We demonstrated through our various empirical performance results that these approaches can achieve improved performance with fewer labelled training instances compared to passive learning, which leads to a savings in annotation efforts for hydroxylation site prediction.

The experimental results also showed that among the active learning strategies, the density-uncertainty-based query strategy, seeded by representative initial training set, showed the greatest performance. This was due to the initial selection of the most representative labelled instances followed by the most informative unlabelled instances for annotation at each iteration. As expected, the certainty-based approach, which selects the unlabelled instances most likely to represent true positive hydroxylation sites, achieved the best performance over the positive test set, but suffered from poor performance on the negative set due to a paucity of negative instances in its training data.

Furthermore, our results indicate the uncertainty-based active learning strategy is an effective way to handle class imbalance among the available unlabelled data. By focusing the learning on unlabelled instances near the decision boundary (hyperplane), more balanced class distributions were provided to the learner in each step of the learning process, resulting in an optimal separating hyperplane. Now that simulations have demonstrated the advantage of uncertainty-based active learning strategy for N/D hydroxylation site prediction, we have now passed all unlabelled asparagine/aspartate instances from all human proteins into the hydroxylation site prediction system. As detailed in Chapter 4, we are now following the active learning procedure and a number of unlabelled instances have been selected for labelling and are currently being validated and verified in wetlab experiments. Once the wetlab experiment (i.e., the Oracle) validations are complete, the data will be added to the training set to rebuild the model and the active learning cycle will continue.

4 CHAPTER: SELECTION OF PUTATIVE N/D HYDROXYLATION SITES FOR WETLAB VALIDATION

Unlike other studies that only simulate active learning, the present chapter provides the results of actually applying one active learning cycle to our problem of developing a prediction system for protein asparagine/aspartate (N/D) hydroxylation. Active learning is an iterative process that first trains the best possible classifier using available training data, then, applies the classifier to all unlabelled data in terms of how informative each data point would be if it were to be labelled and added to the training data. At this stage of the active learning iterative cycle, the user must select which points to validate and actually perform the validation. This requires lab-based experimentation and the production of recombinant proteins to determine if hydroxylation occurs at the predicted sites. Certain wetlab experimental limitations must be considered prior to undertaking protein validation and these are outlined in the thesis. Furthermore, several bioinformatics analyses required to select the subset of proteins that were putatively hydroxylated for the wetlab validation would be described. In this chapter, we briefly describe the actual stage involved in wetlab validation of a putative hydroxylation site on a protein and the results from individual stages of the wetlab validation process are presented.

4.1 Unlabeled dataset collection and data preprocessing

(Liu, 2009) made the 1,288,896 potential N/D hydroxylation sites available. For the dataset, the first 6 columns of Table 4.1 was given but 7^{th} column was computed as detailed in section 4.1.1, where first column is the serial number (S/N); the second column contains protein Accession ID Number, Version Number and Site Location of the asparagine or aspartate on each protein,

column 3 designates whether the site is an N or a D, column 4 contains the labels positive (+1) or negative (-1), column 5 is the hydroxylation status of the proteins written as either "YES" (equivalent to hydroxylated) or "NO" (meaning non-hydroxylated), column 6 is the confidence values of the prediction expressed as a percentage (%), and column 7 is the decision values as shown in Table 4.1. We performed data preprocessing to remove all experimentally verified protein data (1813) from the 1,288,896 and the remaining 1,285,299 N/D sites were taken as the unique unlabelled N/D dataset (i.e., true label of these data points are unknown) used for the active learning cycle.

4.1.1 Preprocessing SVM prediction score data

The trained SVM classifier provided only a binary prediction (Yes/No) and a confidence value (0-100%), however a decision value (score) was required to apply the various active learning query strategies described previously. Algorithm 4.1 was used to compute the decision values and column 7 of Table 4.1 illustrates some sample output.

Table 4.1. Source of unlabeled N/D dataset used for active learning

S/N	Accession_ID	N/D	Prediction	Status	Confidence (%)	Dec_values
1	NP_112234.2_569	N	-1	NO	99.92	0.000395
2	NP_077719.2_2004	D	1	YES	80.49	0.902465
3	NP_004059.2_415	D	-1	NO	0.009	0.499957
4	NP_115823.3_2143	N	1	YES	58.38	0.791915
5	NP_115823.3_2173	N	-1	NO	99.95	0.000234
	⋮	⋮	⋮	⋮	⋮	⋮
1288896	NP_055572.1_414	N	-1	NO	99.84	0.000780

```
if (Yes)
    dec_values = 0.5 + confidence values (%)*(0.5)
else
    dec_values = 0.5 - confidence values (%)*(0.5)
end
```

Figure 4.1. A simple formulation to obtain decision values from a binary classification and a prediction confidence value.

4.1.2 Active Learning query strategies

The SVM pool-based active learning algorithm discussed in Section 3.2 and Algorithm 3.1 were used to assess the informativeness of each of the 1,285,299 putative asparagine/aspartate hydroxylation sites. Active learning was induced with the unannotated asparagine/aspartate hydroxylation sites, using two active learning query strategies: uncertainty-based sampling and certainty-based sampling.

Although the density-uncertainty query strategy outperformed other active and passive learning strategies, it was not considered in the implementation of the active learning prediction for the real unannotated N/D hydroxylation sites since the initial training set was available from a previous study (Liu 2009). Furthermore, it also uses the same query technique as uncertainty-based query; therefore, it is most likely to produce the same result as uncertainty-based strategy. In uncertainty-based query strategy, the active learner selects unannotated hydroxylation sites

considered most informative or most uncertain (i.e. maximum uncertainty) for annotation for the wetlab experimental validation. The selection of such unannotated N/D hydroxylation sites (i.e. site the classifier is least confidence on how to classify) is most likely to influence the decision boundary in the next active learning cycle when added to the training set.

The certainty-based strategy selects those unannotated N/D hydroxylation sites with the strongest confidence of most likely to be annotated via wetlab validation (i.e. the Oracle). This strategy was chosen since we are equally interested in improving the future classification performance and also in confirming novel N/D hydroxylation sites.

During the active learning process, uncertainty-based and certainty-based query strategies were implemented using all labelled data points (1,813) as the initial training sets, and all remaining N/D locations (i.e. 1,285,299 minus 1,813 sites) were used as the pool of unannotated N/D hydroxylation sites. Then, the uncertainty-based active learning query strategy queried the most informative unannotated hydroxylation sites, while the certainty-based active learning query strategy chose unannotated hydroxylation sites with the strongest confidence of being true hydroxylation sites. Both active learning query strategies (certainty-based and uncertainty-based) were used to select the top 20 putative hydroxylation sites on human proteins in terms of desirability to validate its hydroxylation status via wetlab experiments. Once the Oracle validation identifies the true label of the selected putative N/D hydroxylation sites, the newly labelled instances will be added to the training set to rebuild the SVM model and the active learning cycles will continue.

4.1.3 Generation of active learning ranked list of sites

We successfully obtained all 20 protein sequences from the National Centre for Biotechnology Information (NCBI) database to examine and detect the suitability of the tryptic fragments (polypeptides) containing the putative sites for analysis through mass spectrometry (MS) for the determination of the true annotation of these putative N/D hydroxylation sites. As detailed below, five out of the 20 proteins were chosen for wetlab experimental validation after bioinformatics evaluation of the availability of genes with these target proteins, availability of specific antibodies, and the results of simulated tryptic digestion to unveil the tryptic fragments containing the putative N/D hydroxylated sites.

4.2 Wetlab considerations

4.2.1 Availability of gene clones

The selection of top ranked proteins containing putative N/D hydroxylation sites for wetlab validation was based on the genes for the proteins of interest being previously cloned into mammalian expression vectors that can allow the protein to be expressed in mammalian cell lines. Protein overexpression allows for easier recovery of the protein of interest through a single-step purification method known as immunoprecipitation. Overexpression of the protein of interest in human cell lines ensures that all post-translational modifications of the protein will be similar to that found within human cells. The overexpression of recombinant proteins in mammalian cell lines other than human cells (rat, mouse, monkey) cannot guarantee that such modifications will be the same as that found in humans. The cell line of choice is the human embryonic kidney cell 293 (HEK) cell line and this cell line was utilized for validation of hydroxylation in this thesis.

4.2.2 Availability of antibodies

We required antibodies that could detect and bind specifically to the protein of interest with sufficient affinity, a necessary determinant for a successful immunoprecipitation of these target proteins. By searching for the availability of a suitable antibody for each potential protein of interest, we were able to reduce the list of candidate proteins significantly. Between 10 and 20 amino acids is the length of peptide commonly used for the production of antibody (Open BioSystems, 2011).

4.2.3 Suitability of peptides for characterization via mass spectrometry

Peptides longer than 100 amino acids (residues) cannot be directly sequenced using mass spectrometry (MS) (Voet et al., 2008) and must therefore first be cleaved (cut), typically using the protease trypsin (a protein that cuts other proteins), into smaller fragments amenable to MS analysis. We considered the length of the polypeptide fragments, produced by tryptic digestion, surrounding each of the top-ranked N/D amino acids identified by the active learning strategies. Specifically, we examined the tryptic fragment containing the putative N/D hydroxylation site to ensure our final choice of proteins selected for wetlab validation will have adequate length of at least 10 amino acids such that they will ionize favourably and be amenable to characterization via mass spectrometry.

To examine the tryptic fragments surrounding each highly ranked putative N/D hydroxylation site, it was necessary to simulate tryptic digestion of the proteins. In biological systems, trypsin cleaves peptide chains mainly at the carboxyl side of the amino acids lysine (K) or arginine (R), except when either is followed by a proline (P). The widely accepted rule for a trypsin cleavage

(cut) site is [RK] · [^P]. In this notation, [RK] denotes "either R or K" and [^P] denotes "any amino acid other than P". These filtering rules are used in the leading polypeptide analysis tool *peptideCutter* provided by ExPASy (Expert Protein Analysis System) (Rodriguez et al., 2008). However, when the *peptideCutter* tool was applied, we noticed that the "after R or K unless followed by P" rule was occasionally violated. Therefore, a simple python script (trypsin_digest_script) was written to properly simulate the tryptic digestion. Figure 4.2 illustrates a case where *peptideCutter* fails to properly simulate tryptic digestion (missing third cleavage site) whereas trypsin_digest_script adheres to the rules. See Appendix A.6 for the detailed description of this script. A typical example of how a tryptic digest cleaves a polypeptide, using the trypsin specificity rules, is shown in Figure 4.3. Note that the last putative hydroxylation site (i.e. the **N** in the last row) falls within a very short tryptic fragment. This site would therefore not be a good candidate for wetlab validation since its tryptic fragment may not be amenable to analysis via mass spectrometry.

```
LEMIFAK  FDEVQSSGGMILSVCKDK ← peptideCutter
LEMIFAK  FDEVQSSGGMILSVCK  DK ← trypsin_digest_script
```

Figure 4.2. Example of peptide fragments by peptideCutter and trypsin_digest_script.

KXFLVQFGVNV**N**AADSDGWTPLHCAASCNNVQVCKXF

RXLAGGSGLPGSVDVDECSEGTDDCHIDAICQ**N**APKXS

KXFVLGQCIPEDYDVCAEAPCEQQCT**D**NFGRXV

KXSF**DD**EESV**D**G**N**RPSSAASAFKXV

RXA**N**ALKXK

Figure 4.3. Examples of peptide identification by tryptic digestion rule.

The cleavage site location is marked with an "X" and the potentially hydroxylated amino acid is the "bold" N or D. The amino acids underlined "RP" represents a site that trypsin will not cut. Then, the peptides shorter than 10 amino acids may not ionize in mass spectrometry for the reason given in section 4.2.2.

4.3 Final list of proteins selected for wetlab

Below is a brief description of the N/D proteins chosen, along with their crystal structures and hydroxylation sites identified by the active learning algorithms.

4.3.1 Gene: TP53BP2 (apoptosis-stimulating protein of p53 isoform 1 or ASPP2)

TP53BP2 also known as ASPP2 (apoptosis-stimulating of p53 protein 2), is a tumour-suppressing p53-binding protein (Naumovski and Cleary 1996). ASPP2 is one of the three members of ASPP family (along with ASPP1 and iASPP) with the most known protein-binding partners, as shown in Figure 4.4. The N-terminus of ASPP2 has the structure of a β-grasp ubiquitin-like fold (Tidow et al. 2007) while the C-terminal part contains four ankyrin repeats and an SH3 domain involved in protein-protein interaction (Gorina and Pavletich 1996). It is found in the perinuclear region of the cytoplasm, and regulates apoptosis and cell growth, through interactions with other regulatory molecules including members of the p53 family of proteins (Nakagawa et al. 2000; Samuels-Lev et al. 2001). Apoptosis, or programmed cell death, is a major control mechanism by which cells undergo death to control cell proliferation (Lowe and Lin 2000). ASPP1 activates p53, which in turn, causes apoptosis to occur. This is particularly important in the demise of cancer cells which undergo uncontrolled proliferation. Thus p53 is known as an oncogene (a gene that has the potential to cause cancer). The

hydroxylation site identified by certainty-based query strategy on this protein is asparagine (N-985), as shown in Figure 4.5.

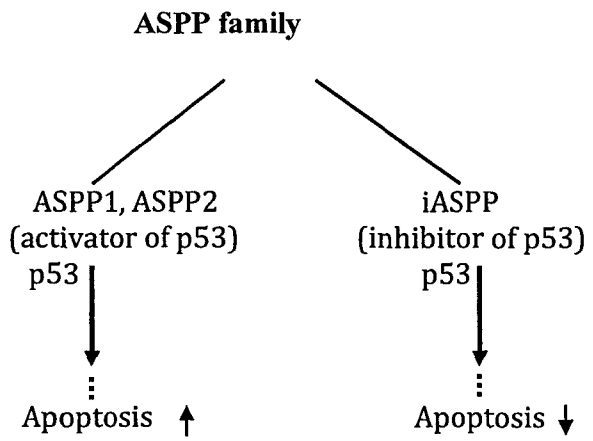

Figure 4.4. ASPP members are apoptotic specific regulators of p53.

The apoptotic function of p53 is stimulated by ASPP1 and ASPP2 and is inhibited by iASPP (Ze-Jun et al., 2005).

4.3.2 Gene: PPP1R13L (RelA-associated inhibitor)

The RelA-associated inhibitor, also known as iASPP, is a protein that is encoded in humans by the gene iASPP. iASPP is the third member of the p53-regulating ASPP family of proteins, the most evolutionarily conserved inhibitors of p53. The C-terminal part of iASPP also contains four ankyrin repeats and an SH3 involved in protein-protein interaction.

The contribution of ankyrin repeat residues, and those of the SH3 domain, generates distinctive architecture at the p53-binding site, suitable for inhibition with small molecules (Robinson et al. 2008). iASPP can block ASPP1 and ASPP2 from binding to p53 and stimulating apoptosis.

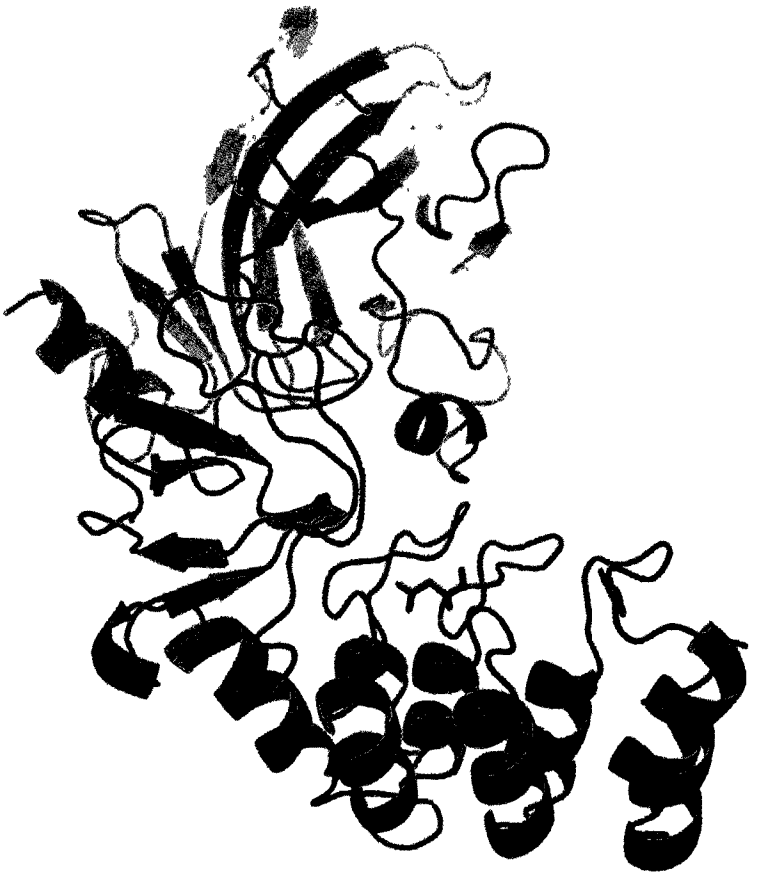

Figure 4.5. The crystal structure of ASPP2 (cyan) bound to p53 (green) from Protein Data Bank (PDB) ID 1YCS (Gorina and Pavletich 1996). The putative hydroxylation site (N-985) on ASPP2 is shown in red. N-985 was identified by the certainty-based query strategy.

Thus, the competition between negative regulation by iASPP and positive regulation by other members of the ASPP family could determine the apoptotic status of p53 and, ultimately, cell fate (Bergamaschi et al., 2003). The PPP1R13L gene is over-expressed in different cancers and its expression in the p53 wild-type background is sufficient to promote tumour growth in vivo. In

an un-transformed (non-cancerous) cell, it acts as a tumour suppressor. However, the modulatory effect of the RelA-associated inhibitor protein on the ability of p53 to cause cellular apoptosis has important implications in cancer and presents new therapeutic possibilities (Laska, et al., 2010). The hydroxylation site on this protein, identified by certainty-based query strategy, is asparagine (N-687), as shown in Figure 4.6.

Figure 4.6. The PPP1R13L crystal structure from PDB ID 2VGE (Robinson et al., 2008). The PPP1R13L is predicted to be hydroxylated on the alpha helix N-687 residue (red) identified by the certainty-based query strategy.

4.3.3 Gene: AP2M1 (AP2 complex, subunit MU, isoform B)

This gene encodes a subunit of the heterotetrameric coat assembly protein complex 2 (AP2), which belongs to the adaptor complex's medium family of subunits (Druck et al., 1996). The

encoded protein is required for the activity of a vacuolar ATPase, which is responsible for proton pumping occurring in the acidification of endosomes and lysosomes. The encoded protein may also play an important role in regulating the intracellular trafficking and function of CTLA-4 (cytotoxin T-lymphocyte Antigen 4). Two transcript variants, encoding different isoforms, have been found for this gene. The hydroxylation site identified by uncertainty-based query strategy on this protein is aspartate (D-413) shown in Figure 4.7.

Figure 4.7. The AP2M1 crystal structure from PDB ID 2VGE (Kittler et al., 2008). AP2M1 is predicted to be hydroxylated at D-413 (red) identified by the uncertainty-based query strategy.

4.3.4 Gene: CCBE1 (collagen- and calcium-binding EGF domain containing protein 1 precursor)

This gene is thought to function in extracellular matrix remodelling and migration. It is mainly expressed in the ovary; but is down-regulated in ovarian cancer cell lines and primary carcinomas, suggesting its role as a tumour suppressor (Barton et al., 2010; Alders et al., 2009).

Mutations in this gene have been associated with Hennekam lymphangiectasia-lymphedema syndrome, a generalized lymphatic dysplasia in humans (Hennekam et al., 1989). A certainty-based query strategy identified a potential hydroxylation site on the protein at D-104. Unfortunately, there is no crystal structure for CCBE1 to date.

4.3.5 Gene: LTBP3 (latent-transforming growth factor b-binding protein, variant 2, or LTBP3)

The protein encoded by this gene forms a complex with transforming growth factor beta (TGF-beta) proteins and may be involved in their subcellular localization. Activation of this complex requires removal of the encoded binding protein. This protein may also play a structural role in the extracellular matrix. Three transcript variants, encoding different isoforms, have been found for this gene. LTBP3 was predicted to be hydroxylated at D-505 by the uncertainty-based query strategy. Unfortunately, no crystal structure is currently available for LTBP3.

4.4 Wetlab experimental validation (Oracle)

Wetlab experimental validation of the aforementioned putative N/D sites, identified by the active learning query strategy, was undertaken. The wetlab experimental validation serves as the Oracle in the second phase of the active learning cycle, which aims to determine the true annotation of the five putative N/D hydroxylation sites considered. Sometimes, the Oracle may not be able to label some N/D data points, therefore, the experimenter would move to the next best point. An overview of the individual steps of the wetlab experimental validation is given in Figure 4.8. These hydroxylation sites, once validated to be positive or negative, would be added to the

training set and the performance of the active learning will be evaluated on an independent test set to complete the active learning cycle.

4.4.1 Amplification and isolation of plasmid DNA

A plasmid is a small double-stranded, circular DNA molecule that replicates independently of the chromosome in bacteria cells. Plasmids are used to carry one or more genes, and must contain an origin of replication, promoters, antibodies resistance genes, a gene of interest to be expressed and a polylinker region. A polylinker (also referred to as a multiple cloning site or MCS) is a region of DNA within a cloning vector that contains various recognition sites for a wide variety of restriction enzymes. Usually the gene of interest is cloned into the plasmid within the polylinker region.

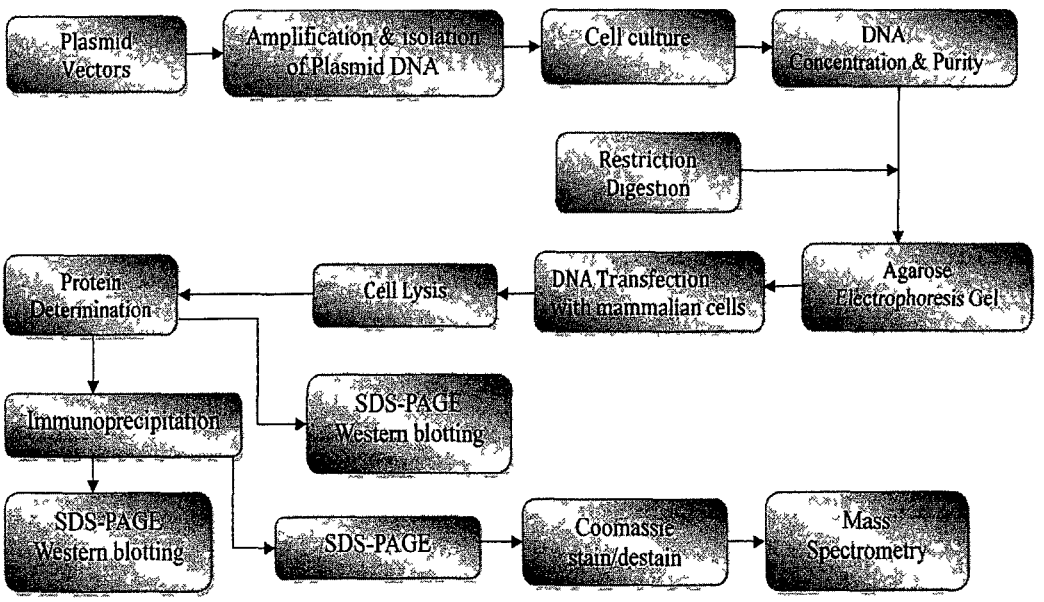

Figure 4.8. Wetlab Experimental Validation Workflow.

The plasmid vector, containing the genes to be expressed as proteins in mammalian cell lines, is known as pCMV-SPORT6 (Figure 4.9). The target genes cloned into their vector included TP53BP2, PPP1R13L, AP2M1 and CCBE1. The fifth target gene, LTBP3, was cloned into a pBluescriptR vector (Figure 4.10) which is not constructed for expression of the target gene in mammalian cells. LTBP3 was therefore subcloned from pBluescriptR into pCMV-SPORT6. Each of these plasmids contains a T7 promoter and sites where restriction enzymes cut, including *NotI*, *EcoRV*, *SalI*, *BamHI* and *XhoI*.

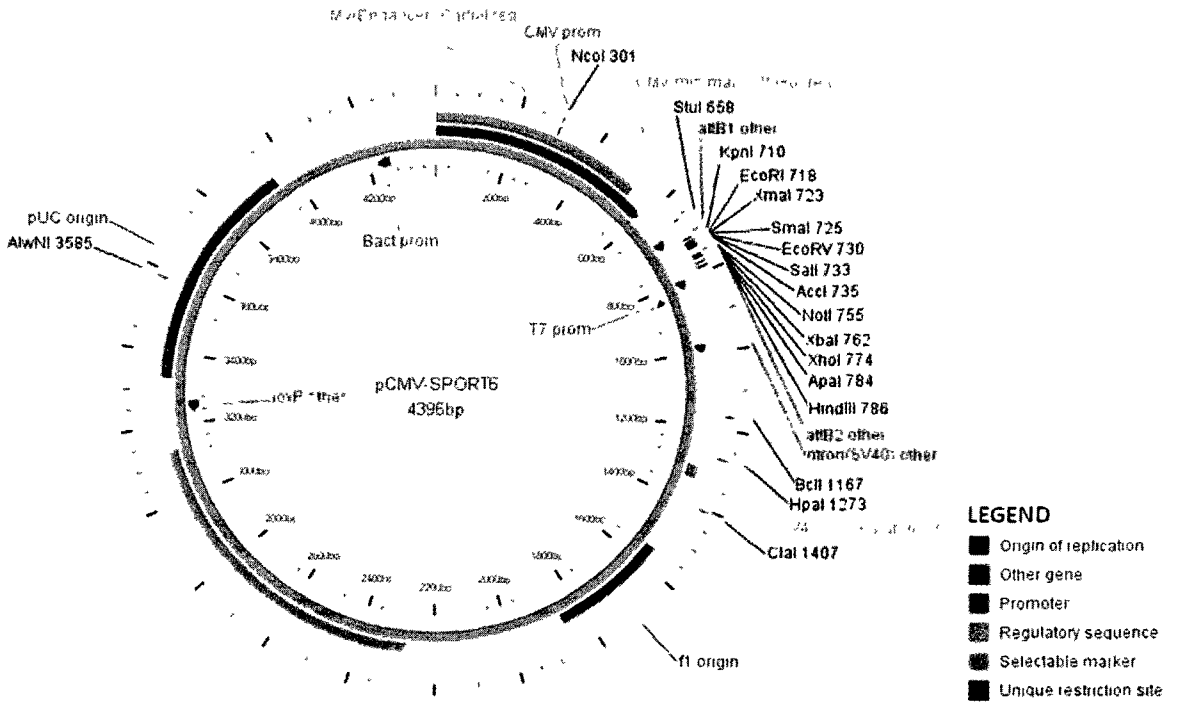

Figure 4.9. The vector map of pCMV-SPORT6 (Dong et al., 2004).

Plasmid DNA (10 ng) was mixed with competent bacterial cells (50 µL) and transformed by heat shocking the cells at 37°C for 45 seconds. Bacteria were then plated out on Luria Bertani (LB) agar plates containing 100 µg/mL ampicillin. Bacteria colonies containing these plasmids were

isolated and grown (which replicates the plasmid) in 100 mL of LB broth containing 100 μg/mL ampicillin (as the plasmids confer ampicillin resistance

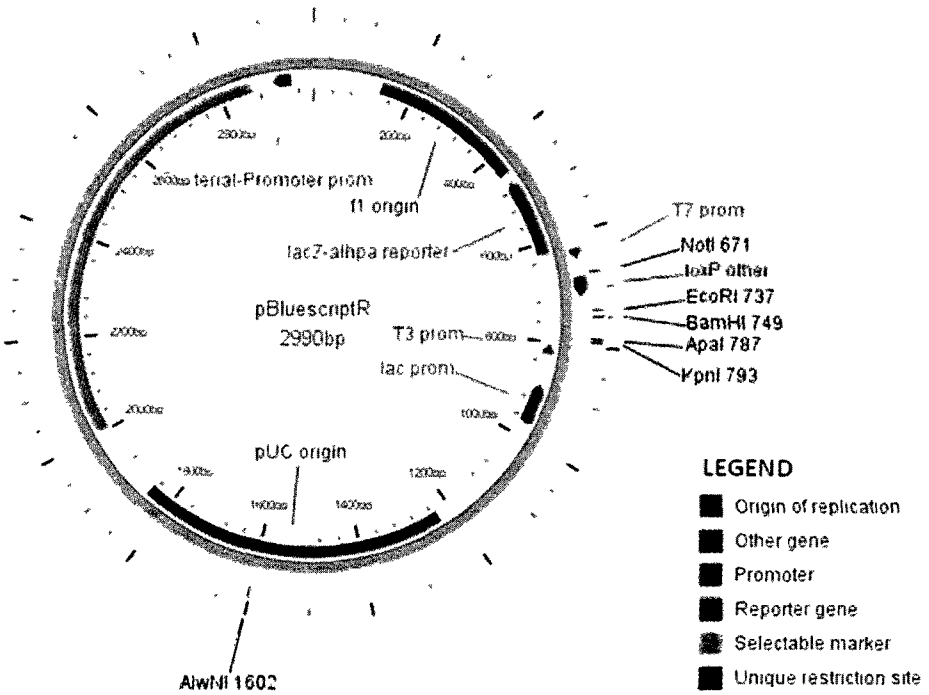

Figure 4.10.The vector map of pBluescriptR. (Dong et al., 2004).

to the bacteria which contain them) overnight with vigorous shaking. Bacteria were harvested the next day by low-speed centrifugation and lysed to obtain plasmids. The five plasmids were purified using a Wizard® Plus Midiprep DNA purification system (Promega, Madison, Wisconsin). Successful plasmid purification was validated through restriction mapping and assessment by DNA agarose gel electrophoresis (see below).

4.5 Wetlab validation techniques and results

4.5.1 Results

Following plasmid isolation, the DNA concentrations were determined by UV spectrophotometry (CARY 100 Bio UV-Visible Spectrophotometer, Varian/Agilent Technologies, Santa Clara, California) using 5 µL of each purified DNA plasmid in 495 µl of Milli-Q water. We estimated the DNA concentration and purity with an absorbance wavelength scan between 230 and 320 nm. The concentration of each DNA sample was then calculated by an absorbance at 260nm (A_{260}), the wavelength at which DNA absorbs light most strongly. The A_{260} was then used to calculate the concentration of DNA according to the following formula:

One absorbance (A_{260}) unit of double-stranded DNA equals 50 µg/mL.

Dilution factor (DF) = total volume of DNA solution/DNA sample volume.

Then, the concentrations (µg/µl) of the purified DNA samples were obtained as follows:

$$[\text{DNA}](\mu g/\mu l) = \frac{Abs_{260} \text{ x } 50\ \mu g/ml \text{ x } 100(DF)}{1000\ \mu l/ml}$$

4.5.2 Restriction enzyme digestion and DNA gel electrophoresis

Purified plasmid DNA was subjected to digestion by restriction enzymes to confirm successful isolation. Restriction digestion was performed using restriction enzymes BamHI, SalI, XhoI, NotI and EcoRV. Digested DNA plasmids were further analyzed using agarose gel electrophoresis. An agarose gel electrophoresis is a technique used to separate DNA, or RNA molecules by size. The separation of the DNA molecules is achieved by subjecting the gel to an electric field where negatively charged DNA migrates through the agarose gel matrix toward

ethylenediaminetetraacetic acid, pH 8.0). An intercalating DNA dye (Red Safe, FroggaBio, Toronto, Canada) was added to the gel to visualize DNA and a 1 Kb DNA ladder (molecular weight standard) was used to determine the molecular weight of the resulting digestion bands. Samples of the purified DNA plasmids were loaded into wells of the agarose gel as shown in Figure 4.12 and exposed electric field for 30 minutes at 150 volts. The migration of the DNA through the gel is then visualized under UV illumination and the gel was imaged using an Alphaimager gel documentation system (ProteinSimple, Santa Clara, California). The size of the restriction fragments were estimated according to the DNA ladder. The gel electrophoresis procedure was performed according to Promega Protocols and Applications Guide (2011).

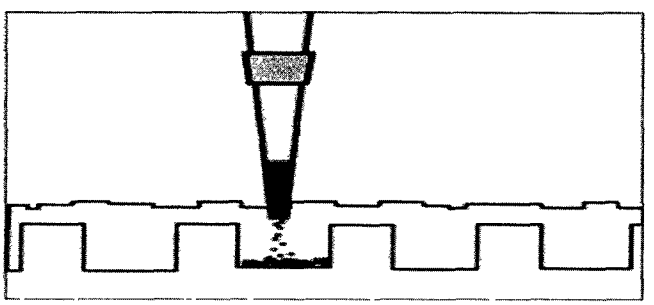

Figure 4.12. Technique of loading DNA samples into a well (adapted from nwabr.org). The pipette tip is placed just above the well before releasing the DNA sample into the well

Results

Restriction digestion of plasmid DNA confirmed successful isolation of the plasmid and the presence of the insert in the vector. Figure 4.13 shows the agarose gel where lanes 1 and 10 correspond to the 1 Kb DNA ladder and lanes 2, 4, 6 and 8 show the undigested plasmids

(controls) of CCBE1, TP53BP2, AP2M1 and PPP1R13L respectively. Lane 3 corresponds to the CCBE1 plasmid digested with NotI and EcoRV restriction enzyme. Restriction digestion of TP53BP2 with NotI and SalI in Lane 5 resulted in a single band approximately 4,100 bp DNA fragment) of the correct molecular weight for the gene inserted into the plasmid (as the vector and insert were of the same size). Restriction digestion of AP2M1 with NotI and SalI, shown in Lane 7, also generated two bands (approximately 4,580 and 1,867 bp DNA fragments for vector and insert respectively), showing the correct molecular weight for DNA fragments. Restriction digestion of PPP1R13L with NotI and SalI, shown in Lane 9, did not show the correct banding pattern and thus this clone (gene) was not taken to further testing.

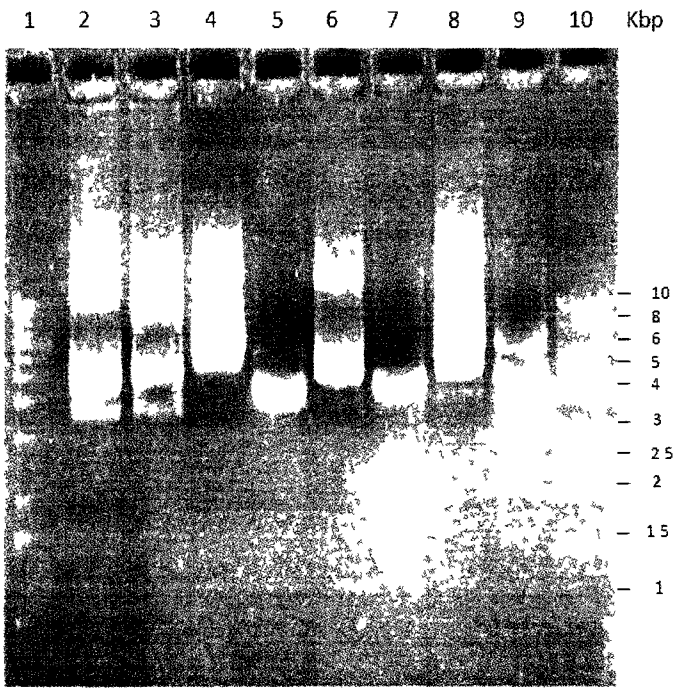

Figure 4.13. DNA fragmentation of DNA plasmids (CCBE1, TP53BP2, AP2M1 and PPP1R13L) in a 1% agarose gel treated with restriction digestion enzymes (BamHI, SalI, XhoI, NotI and EcoRV).

Lanes 1 and 10, 1Kb DNA ladder; Lane 2, undigested CCBE1 control; Lane 3, CCBE1 digested with NotI+EcoRV, 4 undigested TP53BP2 control; Lane 5, TP53BP2 digested with NotI+SalI; Lane 6 undigested AP2M1 control; Lane 7, AP2M1 digested with NotI+SalI; Lane 8 undigested PPP1R13L control; Lane 9, PPP1R13L digested with NotI+SalI.

Initial restriction digests of the plasmid containing LTBP2 did not reveal the expected banding pattern and was taken for further restriction digestion testing to determine presence of the insert and orientation of the insert in the vector. Figure 4.14 shows the 1 Kb DNA ladder in Lanes 1 and 9 and the undigested plasmid (control) in Lane 2. The LTBP2 plasmid digested with different restriction enzymes (BamHI, SalI, and XhoI), shown in lanes 3 to 8, did not show the correct fragmentation pattern. The BamHI and SalI digests (lanes 3, 4 and 6) did not reveal the correct fragmentation pattern. Digestion with XhoI (Lane 5) cut the plasmid twice when only one cut was predicted (according to the DNA sequence). Lanes 7 and 8, which were the double (BamHI+XhoI) and triple digests BamHI+SalI+XhoI) respectively, did not show the predicted pattern of DNA fragmentation. Therefore, LTBP2 was abandoned as a potential candidate gene for further testing.

4.5.3 An overview of transient transfection of plasmid DNA into HEK 293 cells using Lipofectamine™2000

Transient transfection is a process of introducing a purified plasmid DNA into mammalian cells, such as human embryonic kidney (293) cells, with an efficient cationic liposome-based reagent (Lipofectamine 2000) and resulting cationic liposome (or mixture) fuses with the negatively charged cell membrane and the DNA is released into the cell.

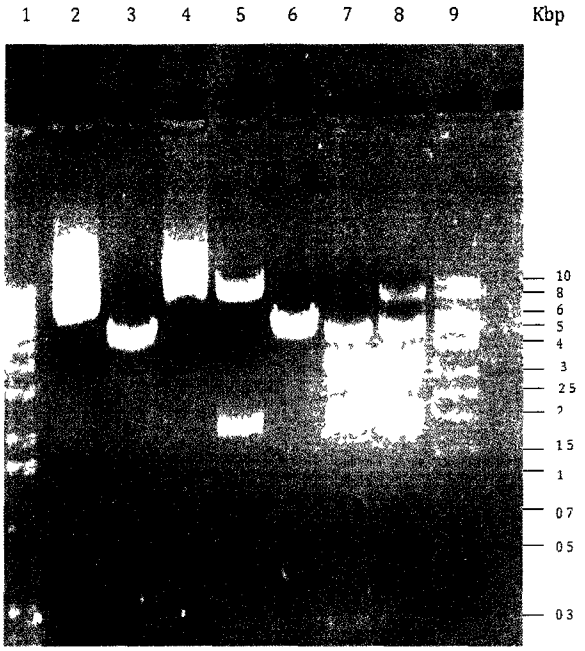

Figure 4.14. DNA fragmentation of plasmid (LTBP2) in a 1% agarose gel with restriction enzymes (BamHI, SalI, XhoI, NotI and EcoRV).

Lanes 1 and 9, 1 Kb DNA ladder; Lane 2, undigested LTBP2 control; Lane 3, LTBP2 digested with BamHI; Lane 4, LTBP2 digested with SalI; Lane 5, LTBP2 digested with XhoI; Lane 6, LTBP2 digested with BamHI+SalI; Lane 7, LTBP2 digested with BamHI+XhoI; Lane 8, LTBP2 digested with BamHI+Sal1+Xho1.

Method

One day prior to transfection, cells were trypsinized and counted to ensure that cell density is 90-95% confluent on the day of transfection. We used a liposome-based mediated transfection method (Lipofectamine 2000 reagent) to introduce the various DNA plasmids into cultured HEK 293 cells. This transfection method involves mixing the DNA plasmid to be transfected with

Lipofectamine 2000 reagent. We followed the manufacturer's protocols on transfection of HEK293 cells and a detailed step-by-step of the transfection method is shown in Figure 4.15. To optimize transfections, the amount of DNA added to the mixture was varied from 2 to 4 μg for each DNA plasmid, in order to determine the optimal concentration of DNA for protein expression in mammalian cells.

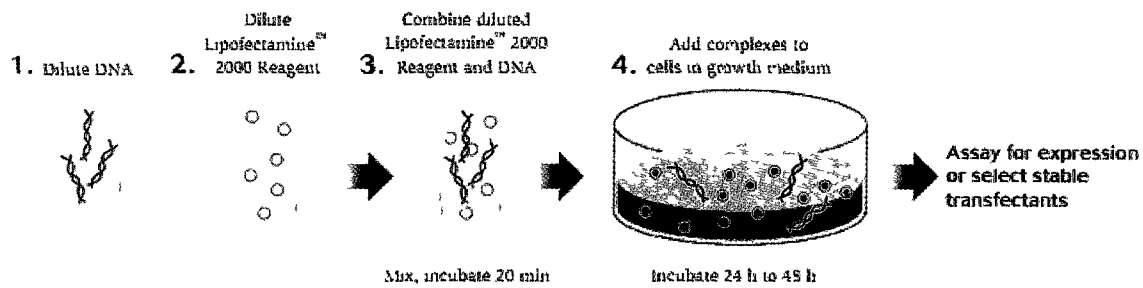

Figure 4.15. Outline of the transient transfection procedure for plasmid DNA into HEK 293 cells using Lipofectamine™ 2000 Reagent (Invitrogen, Carlsbad, California).

In stage 1, 2 and 4 μg of each plasmid DNA was diluted in 250 μL of Opti-MEM I (Reduced Serum Medium); Stage 2, 10 μL of Lipofectamine 2000 was diluted in 240 μL of Opti-MEM I (Reduced Serum Medium); Stage 3, the diluted DNA solution was combined with diluted Lipofectamine 2000 and incubated at room temperature for 20 minutes to allow DNA-Lipofectamine 2000 liposomes complexes to form; Stage 4, Lipofectamine 2000 and plasmid DNA complex mixtures were added to the HEK293 cells in 6-well plates and incubated 37°C for 24 hours. The relative surface area and working volume for the 6-well plates can be found in Appendix A.3.

Results

After 48 hours of transfection, cells were harvested using cell lifters and placed in 15 mL Falcon tubes and centrifuged at 5,000 rpm for 2 minutes at 4°C in a Thermo IEC MultiRF centrifuge (Thermo Fisher Scientific, Waltham, Massachusetts). The supernatant removed, and the cells flash frozen in liquid nitrogen and stored at -80°C for future use.

4.5.4 Lysis of transfected HEK 293 cells

An optimized cell lysis was also performed for normoxic and hypoxic treated HEK293 cells, transfected with AP2M1 and TP53BP2. Cell lysis involved disrupting the cell membranes in order to release the protein of interest. Lysis action solubilizes the proteins in both the cell cytoplasm and nucleus so they can migrate through a separating polyacrylamide gel. Pelleted cells were removed from -80°C freezer, thawed and 1 mL of lysis buffer (20 mM HEPES, pH 7.9, 420 mM NaCl, 0.2 mM EDTA, 25% glycerol, 1.5 mM MgCl2, Roche Complete Protease Inhibitor Tablet) was added to each tube containing transfected HEK 293 cells and incubated on ice for 20 minutes. Cells were then centrifuged for 15 minutes at 13,000 rpm at 4oC. The resulting supernatant was transferred to a new Eppendorf tube.

4.5.5 Bio-Rad Assay Protein Concentration Determination

The Bio-Rad Protein Assay Dye Reagent Concentrate is a colorimetric assay for determining protein concentrations of samples. A standard curve, utilizing bovine serum albumin as a standard protein, must be prepared every time the assay is performed.

The protein concentrations of cell lysates were determined using a Bio-Rad protein assay. A series of bovine serum albumin (BSA) standards were prepared with the following concentrations: 0, 0.05, 0.1, 0.2, 0.4, 0.5, 0.8, 1, 2, 3 μg of protein per well of a 96-well microplate. The Bio-Rad Protein Assay Dye Reagent Concentrate was diluted 1 part concentrate to 4 parts sterile water (i.e. 1:4 diluted Bio-Rad Dye Reagent). Cell lysates from normoxic and hypoxic transiently transfected (AP2M1 and TP53BP2) HEK293 were diluted to 1:10 and 1:20, and 20μl of each solution (cell lysate) and BSA standard curve were transferred into separate wells of a 96-well microplate (in duplicate). Subsequently, 180 μL of 1:4 diluted BioRad Dye Reagent was added, mixed thoroughly and incubated for 5 minutes at room temperature. The absorbance at 595 nm (Abs_{595}) was determined using the Molecular Devices Spectramax 340^{PC} microplate reader (Molecular Devices, Sunnyvale, California) and a standard curve was plotted of absorbance versus the concentrations of BSA.

Using the standard curve, we then determined the protein concentration of each cell lysate from its absorbance values by interpolation (i.e. equation of the line, $y = A + Bx$) given by the BSA standard curve, where y corresponds to the absorbance values at 595nm (Abs_{595}), $A = 0.014$, $B = 0.152$ and x corresponds to the initial protein concentration in μg.

$$\mathbf{Abs_{595}} = \mathbf{0.014 + 0.152}x, \quad [x] = \frac{\mathbf{Abs_{595} - 0.014}}{\mathbf{0.152}}\ \boldsymbol{\mu g}$$

[Protein concentration] (μg/ μl) = x (μg) x DF (dilution factor)/ volume of protein (μl)

Resultant protein concentration gives an estimation of how much protein is needed to run western blotting or immunoprecipitation.

4.5.6 Overview of immunoprecipitation

Immunoprecipitation (IP) is a biological extraction method of precipitating a protein of interest out of cell lysates solution using specific antibodies that specifically binds to that particular protein.

Methods

We prepared a 125 μg protein sample in a new Eppendorf tube from both normoxic and hypoxic supernatant (cell lysates). Two μg of polycolonal and monoclonal antibodies were then added to the cells lysates in each tube, and the cell lysates/antibodies mixture was incubated by gently rocking on an orbital shaker overnight at 4°C. We captured the immunocomplexes (samples) by adding 20 μL of protein A/G agarose bead slurry (Santa Cruz Biotechnologies, Santa Cruz, California). Samples were then centrifuged at 1,000 g for 5 minutes at 4°C to pellet the beads and supernatant was discarded. Pellets were washed four times in 1 mL of Phosphate Buffered Saline (PBS), the centrifugation repeated (1,000 g for 5 minutes at 4°C) and the supernatant discarded. The pellets were then resuspended in 25 μL of 2X Laemmli buffer and boiled for 3 minutes at 95°C to dissociate the immunocomplexes from the beads and loaded immediately onto sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE).

Results

SDS-PAGE was performed and the resulting gel was either stained in Coomassie Blue G-250 or Western blotted onto polyvinyldifluoride (PVDF) membrane.

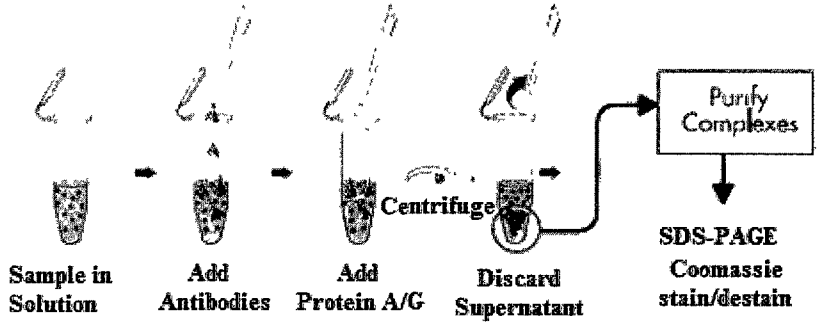

Figure 4.16. Immunoprecipitation workflow where protein complexes are bound to an antibody against a specific protein (reproduced from http://www.millipore.com/immunodetection/id3/immunoprecipitation).

The antibody, in turn, is bound by the protein A/G beads and the entire complex is brought down by centrifugation

4.5.7 Overview of Sodium Dodecyl Sulfate (SDS)-polyacrylamide Gel Electrophoresis (PAGE)

The SDS-PAGE technique is used with reducing agent and detergent (SDS) to separate protein molecules according to their unique molecular weight/size in a porous gel (polyacrylamide) matrix under an electric field. The SDS binds and denatures the proteins to their constituent polypeptides and surrounds the polypeptides with negative charge, such that, migration through the gel is proportional to the molecular weight of the polypeptide.

Methods

SDS-PAGE is composed of a 12% resolving gel and 4% stacking gel (Table A.2). Sixty μg of cell lysate protein was mixed 1:1 with and equal volume of 95% Laemmli buffer and 5% β-mercaptoethanol and separated by SDS-PAGE, see Appendix A.4.

Results

SDS-PAGE gel with protein bands were separated according to their molecular size.

4.5.8 Western blotting Technique

Western blotting is a technique to detect target proteins with specific antibodies, after these proteins have been separated by an SDS-PAGE gel electrophoresis according to their molecular size. The gel is placed next to the PVDF membrane, and then application of electrical current enables the proteins transfer onto the membrane from the gel. This membrane is an exact copy of the gel's protein pattern.

Method

The whole cell lysates were separated by the SDS-PAGE was transferred to a membrane in 20mL of transfer buffer, loaded into a Western blot apparatus for the Western blot process, and the detailed protocol of Western blot can be found in Appendix A.4

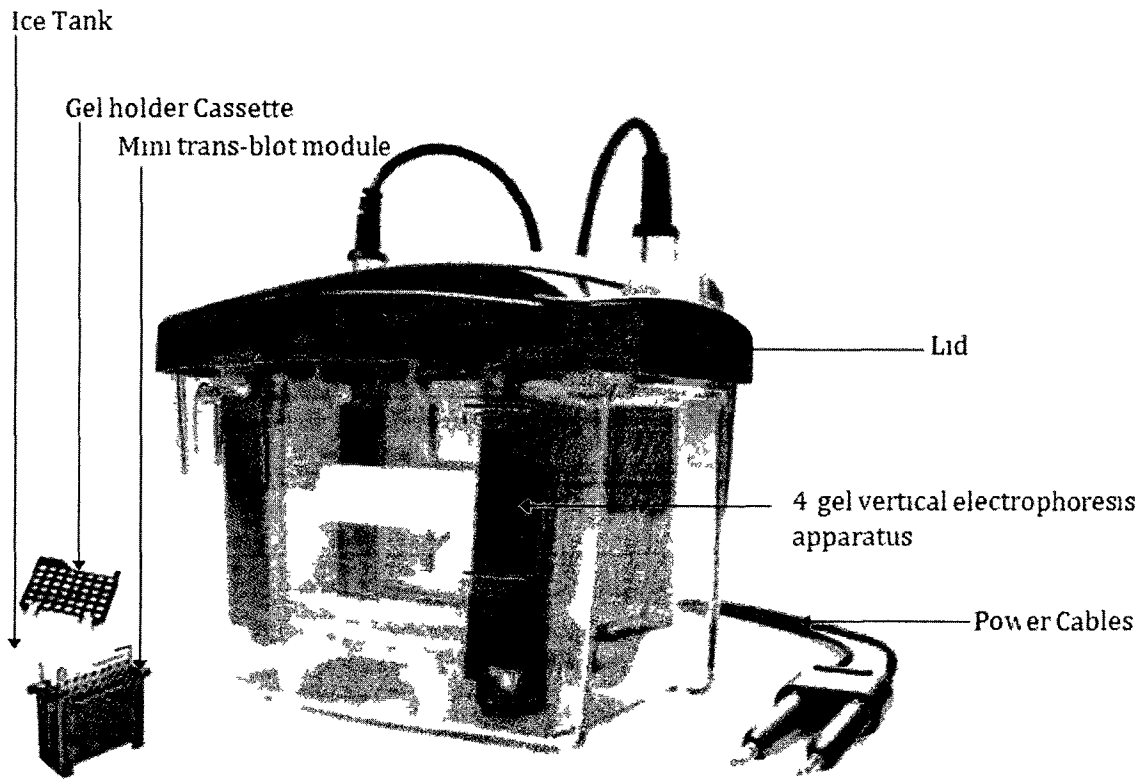

Figure 4.17. A vertical electrophoresis apparatus for slab gel analysis (reproduced from bio-rad.com)

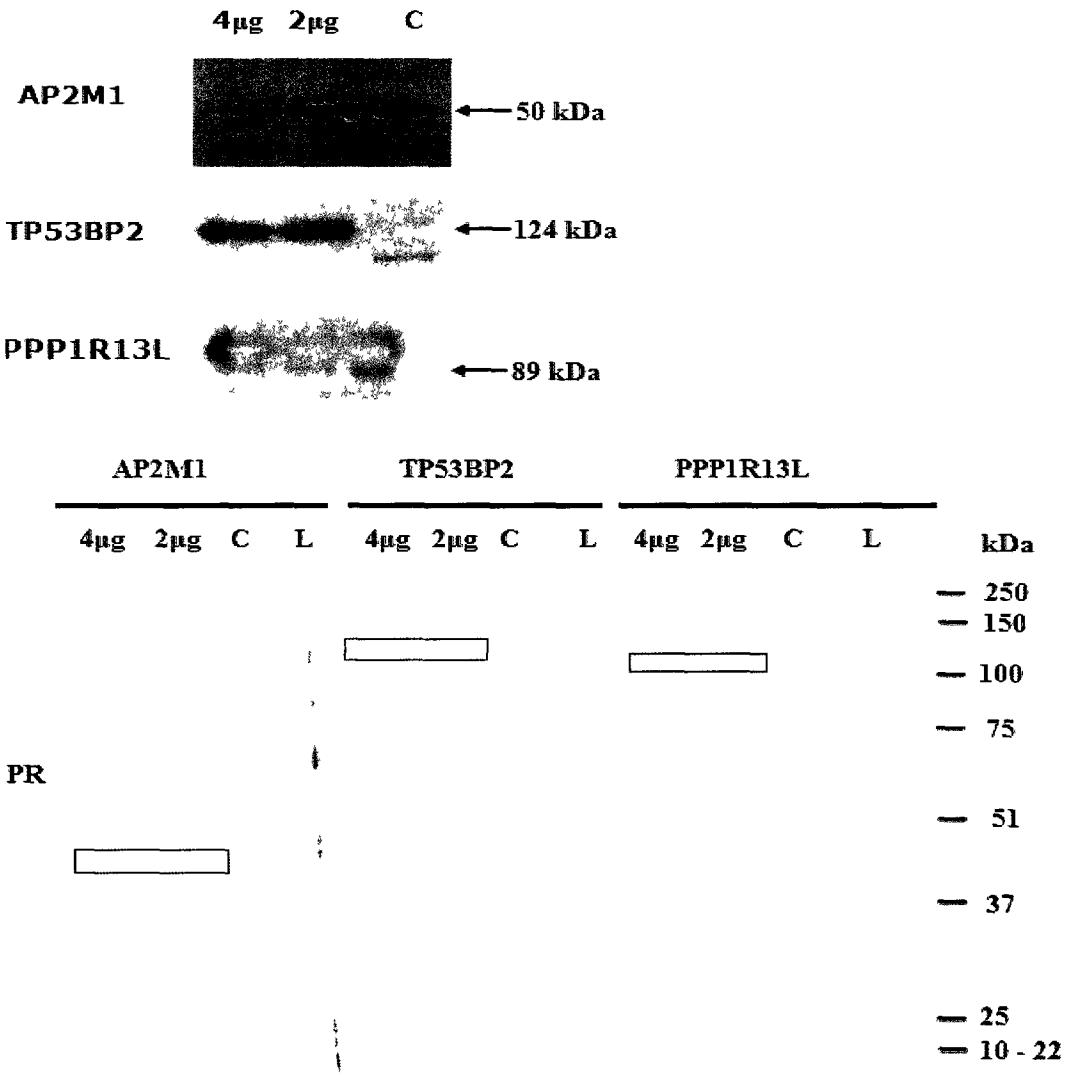

Figure 4.18 Results from a Western blot.

A whole cells lysate from HEK 293 cells transiently transfected with plasmids overexpressing 4µg or 2µg amount for AP2M1 (first panel) or TP53BP2 (second) or PPP1R13L (third panel) or untransfected control (C) treated in normoxic condition were run on SDS-PAGE. These lysates were run on Western Blot for the overexpression of the individual proteins showed as a bands based on their molecular weights. The western blot stained with Ponceau Red (fourth panel), each

overexpressed target proteins are marked with red squares and standard molecular marker (ladder, L).

In Figure 4.18, the membrane blots of the whole cells lysate of the plasmids (AP2M1 or TP53PB2 or PPP1R13L) transiently transfected with HEK 293 cells and were probed with primary antibodies, anti-AP2M1 or anti-TP53BP2 or anti-iASPP at dilution 1:1000 and horseradish peroxidase (HRP)-conjugted goat anti-Rabbit secondary antibody at 1:2,000 dilution. The bands on the Western blot membranes showed various proteins overexpression (recombinant, from the plasmid) and endogenous protein within HEK 293 cells. Then, the bands observed at 124 kDa for TP53BP2 and 50 kDa for AP2M1 proteins are close and consistent with the expected or observed bands at 126kDa and 50 kDa using both anti-AP2M1 or anti-TP53BP2 antibodies (AbCam, Cambridge, Massachusetts) to identified protein of interest. These proteins seemed to have a better overexpression at a lower amount of the plasmids transfected with the HEK 293 cells. On the other hand, the bands for PPP1R13L showed the PPP1R13L protein expression was not detected by the anti-iASPP and secondary antibody, perhaps due to oversaturated protein expression for the treatment with HEK 293 cells.

4.5.9 Coomassie Brilliant Blue R-250 staining and destaining

Coomassie Brilliant Blue staining is used to detect proteins in the range of 100 to 1000 ng in SDS-PAGE gels. Gels are fixed and then stained in Coomassie blue dye solution, where the dye molecules bind to the protein within the gel to form protein-dye complexes, held together by the combination of van der Waals forces and electrostatic interactions. Gels are then destained to visualize the protein bands on the gel.

Methods

To further determine if AP2M1 (A) or TP53BP2 (T) plasmid can be overexpressed in HEK 293 cells, the genes encoding for AP2M1 (A) or TP53BP2 (T) proteins were transiently transfected and expressed in HEK 293 cells under normoxic (N) or hypoxic (H) conditions. After transfection, cells were lysed in cell lysis buffer (20 mM HEPES (pH 7.9), 420 mM NaCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA and 25% glycerol) and the lysate mixed with anti-AP2M1 (A) or anti-TP53BP2 (T) antibodies and protein A/G beads for immunoprecipitation. Cellular proteins were incubated with antibodies and beads for 30 minutes at 4°C with gentle rotation. Immunoprecipitation involved centrifuging beads at 1,000 x g for 5 minutes at 4°C and washing the beads with ice-cold lysis buffer four times with centrifugation in between. After the final wash, beads were resuspended in approximately 20 µL of cell lysis buffer and the entire sample was boiled and loaded onto a 12% SDS-PAGE gel. SDS-PAGE was run at 150V for one hour and the gel was removed, divided in two, and half was used for Western blotting while the other half was used for Coomassie blue staining. Following electrophoresis, the gel was placed in a fix solution (10% acetic acid, 25% methanol). The gel was incubated overnight in the staining solution. The gel was destained in 10% acetic acid, 25% methanol until the background became transparent and the bands were visible. The Coomassie blue detailed protocol can be found in the Appendix A.5.

Results

The results in Figure 4.19 show that AP2M1 protein was detected by the anti-AP2M1 antibody and that the protein was expressed endogenously in cells under both normoxic and hypoxic conditions and that overexpression of the protein resulted in slightly more protein expression.

Western blot analysis showed the molecular weight of the overexpressed AP2M1to be 50 kDa, although the binding of antibody to nonspecific proteins and/or other components was also observed.

TP53BP2 protein expression was seen under both normoxic and hypoxic conditions without overexpression. An extra band was seen with TP53BP2 with the overexpression of AP2M1. This would suggest that overexpression of AP2M1 had an effect on the endogenous expression of TP53BP2. The expression of endogenous TP53BP2 was greater with hypoxia when AP2M1 was overexpressed but lower in hypoxia whereas overexpression of TP53BP2 was lower in hypoxia (without AP2M1 overexpression).The TP53BP2 protein occurred at the correct molecular weight of 124kDa.

Ponceau red stained blots (Figure 4.19) showed two specific bands (show with asterisks) that came down with the beads.

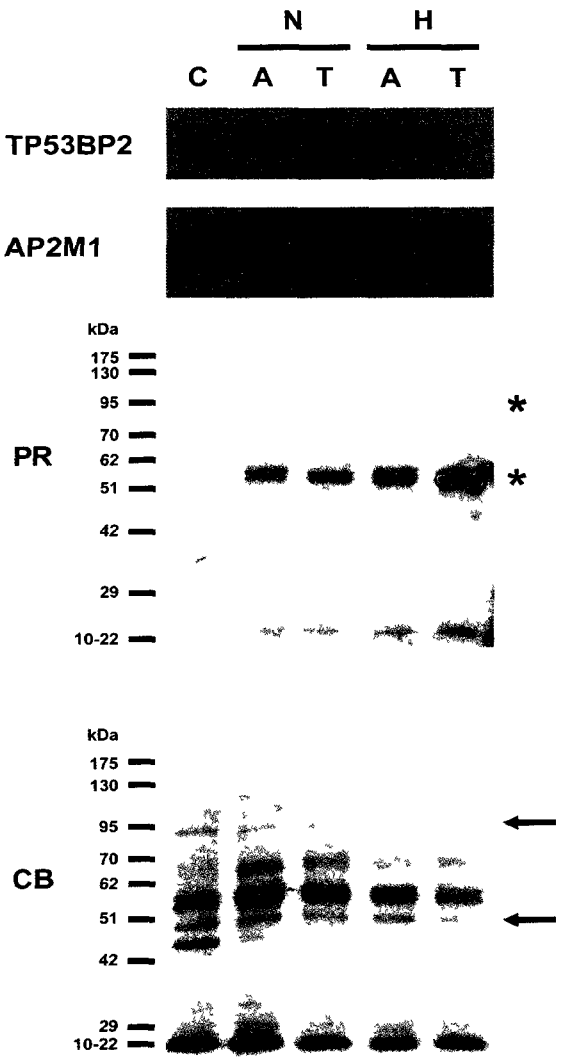

Figure 4.19.Results from an immunoprecipitation on SDS-PAGE gels.

Immunoprecipitations of whole cell lysates from HEK293 cells transfected with plasmids overexpressing either AP2M1 (A) or TB53BP2 (T) or untransfected controls (C) and treated with either normoxic (N) or hypoxic (H) conditions. Lysates were either immunoprecipitated with either anti-AP2M1 (A) or anti-TB53BP2 (T) or protein A/G agarose beads. The Western blots for AP2M1 (first panel) and TB53BP2 (second panel) were performed. Bands that appear in untransfected lanes represent the endogenous protein within HEK293 cells. The immunoprecipitates were run on SDS-PAGE and either Western blotted or stained with Coomassie Blue

(CB; fourth panel). Western blot membranes were stained with Ponceau Red (PR, third panel). Bands from immunoprecipitates that appeared on PR are shown with asterisks. Bands on CB that were taken for subsequent mass spectrometry are shown with arrows.

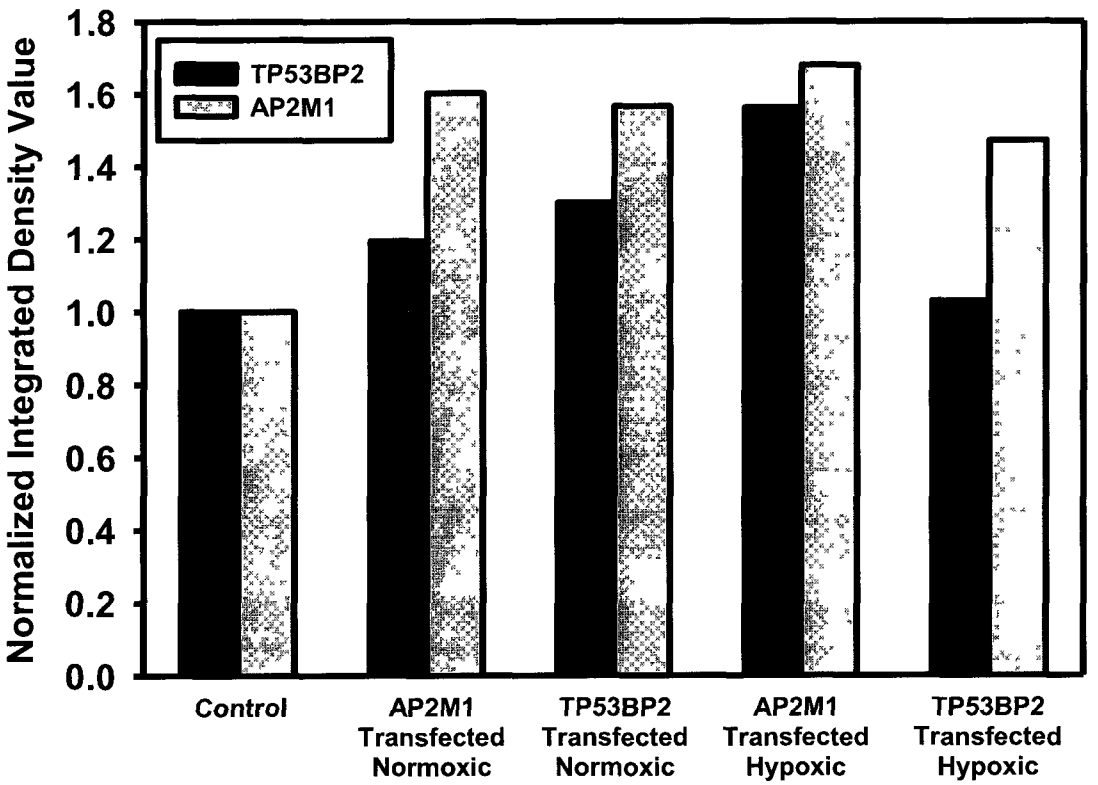

Figure 4.20. Densitometry of bands from Western blots of cell lysates from HEK293 cells. Here, cells have been transiently transfected with plasmids overexpressing either AP2M1 or TB53BP2 or untransfected control cells, and treated with either normoxic or hypoxic conditions. Lysates were either immunoprecipitated with either anti-AP2M1 or anti-TB53BP2 and protein A/G agarose beads.

Figure 4.20 shows the densitometry results (i.e., the integrated density value (IDV)) of each band on the Western blot for AP2M1 or TP53BP2 or untransfected control cells with normoxic or hypoxic conditions; compared to each lane of the Ponceau Red (PR). Bands in Western blots were normalized to the density of a single band on the Coomassie stained gel that did not change with treatment. The highest expression of AP2M1 and TP53BP2 occurred under hypoxic conditions when only AP2M1 was overexpressed. Overexpression of AP2M1 may have stabilized TP53BP2, and this was seen to be enhanced by hypoxia. Overexpressed TP53BP2, however, was destabilized by hypoxia, even under the control of a strong mammalian promoter. The meaning of these changes in protein expression remain to be elucidated.

Specific bands were excised from the gel (indicated by arrows on the Coomassie Blue stained gel), transferred to Eppendorf tubes and stored at -80°C for further mass spectrometry analysis.

4.5.10 Summary

We applied uncertainty-based and certainty-based active learning query strategy to intelligently identified 20 potential N/D hydroxylated sites from 1.3 million putative hydroxylation sites. Five out of the 20 proteins were chosen for wetlab experimental validation after bioinformatics careful analysis. Four (4) of the target genes (TP53BP2, PPP1R13L, AP2M1 and CCBE1) were successfully isolated from plasmid vector, pCMV-SPORT6 and expressed as proteins in mammalian cell lines, human embryonic cells 293 (HEK). The fifth target gene, LTBP2 was successfully subcloned from pBlueSCriptR into pCMV-SPORT6.

All four plasmids were successfully transfected with HEK 293 cells, then, SDS-PAGE and Western blotting were performed to detect the overexpression levels of AP2M2, TP53BP2, and

PPP1R13L proteins in HEK 293 cells using anti-AP2M1, anti-TP53BP2 and anti-iASPP antibodies. The antibody for CCEB1 is not commercially available; as such no further test was conducted for this target protein. Results from Western blotting technique showed TP53BP2 and AP2M1 protein of interest were successfully detected and expressed in HEK 293. However, iASPP antibody could not detect this PPP1R13 protein most likely due to the oversaturated expression, and other reasons were not investigated.

Furthermore, TP53BP2 and AP2M1 proteins were later overexpressed in mammalian cells in normoxic and hypoxic conditions. These proteins biological activity was verified using Western blotting, immunoprecipitation and Coomassie stain analysis based on the overexpression bands identified on an SDS-PAGE gel. These specific bands were excised from the gel, transferred to Eppendorf tubes and stored at -80°C. The successful identification of these bands on the gel lays the foundations for the determination of the true annotation of these putative N/D hydroxylation sites via mass spectrometry.

5 CHAPTER: THESIS SUMMARY

We have demonstrated, through simulations, the applicability of SVM-based active learning in the task of N/D hydroxylation sites prediction. We have also demonstrated that a pool-based active learning query strategy with SVM is able to reduce the effort of protein annotation based on the simulation results for a wetlab experimental validation. The classifier in active learning has the freedom to choose the most informative unlabelled instances and reduces the risk of annotating less useful data points. We successfully overcame class imbalance in the training dataset by selecting instances within the SVM margin surrounding the decision hyperplane. We conclusively reaffirmed the usefulness of margin-based active learning as it able to show when the less informative instances are outside the margin band. We have passed all unlabelled N/D sites dataset from all human proteins into the hydroxylation sites prediction system. Active learning has intelligently chosen the most informative 20 putative N/D hydroxylation sites from 1.3 million putative N/D hydroxylation sites dataset. From this simulated result, we have demonstrated that active learning can indeed reduce annotation for wetlab experimental validation and the wetlab experimental validation component was performed for two proteins that will most likely be amenable to final analysis via mass spectrometry.

It worth mentioning that; the evaluation performance results reported in thesis cannot be directly compared to earlier work by (Liu, 2009) in terms of evaluation metrics such sensitivity (recall), precision rate, Matthew's CC; since we do not use the same learning methods or train/test splits, despite using the same N/D hydroxylation sites dataset. Interestingly, the active learning query strategies identified a number of EGF domain-containing and ankyrin repeat domain-containing

proteins when applied to a real-life N/D hydroxylation sites dataset; this was consistent with Liu's observations (Liu, 2009).

5.1 Summary of contributions

In this thesis, a comprehensive comparison was performed between uncertainty query, density-uncertainty-based query and certainty-based query strategies in terms of prediction recall, precision, MCC, and AUC with growing training set size. We considered the default passive learning approach as the benchmark throughout this thesis. An extensive performance comparison of these methods relative to the N/D protein hydroxylation dataset is provided. We have clearly demonstrated that active learning cycle can drastically reduce the amount of annotation efforts required to obtain a given level of precision and recall for the prediction of hydroxylation sites. Additionally, active learning query strategy handled class imbalance among the available unlabelled data effectively. To the best of our knowledge, a comprehensive comparison between various active learning query strategies for the prediction of a protein post-translational modification has not been reported in previous literature. Our results from the simulation of active learning query strategies for the prediction of N/D hydroxylation sites on human proteins have been accepted for publication in the proceedings of the International Conference on Computational Intelligence and Bioinformatics (CIB 2011), Nov 7-9 2011 in Pittsburgh, PA.

We successfully implemented an active learning query strategy (i.e., uncertainty-based and certainty-based) to intelligently identified the most informative 20 putative N/D hydroxylation sites from 1.3 million putative N/D hydroxylation sites of a real-life dataset. We have selectively

chosen 5 proteins identified by these strategies through the application of bioinformatics analysis to account for wetlab considerations. Two of these proteins were successfully isolated, quantified, over-expressed in mammalian HEK293 cells in an *in vitro* experiment. Then, these proteins' biological activity was verified using Western blotting, immunoprecipitation and Coomassie stain analysis based on the bands identified on an SDS-PAGE gel. The successful identification of these bands on the gel lays the foundations for the determination of the true annotation of these putative N/D hydroxylation sites via mass spectrometry.

5.2 Recommendations for future work

The empirical and wetlab experimental validation component results of this thesis show several interesting future directions. First and foremost is to determine the true annotation of the two putative N/D hydroxylation sites via mass spectrometry. Once validated to be positive or negative, these newly labelled instances should be added to the training set and the performance will be evaluated on an independent test set to complete the active learning cycle. Pursuing the other 3 proteins for which genes have been purchased is also recommended in hopes of further increasing the size of the labelled training set.

Other work we may consider in the future is to propose and implement an active learning stopping criterion that would be dependent on the variance of the classifier's confidence score only for selected unlabelled instances at each iteration. The variance graph can be computed as a function of selected labelled or training set size. It is proposed that an ideal stopping criterion may be when the variance reaches maximum performance and then decreases after subsequent

iterations. Such an implementation could provide a quantitative criterion to determine when to stop learning without human supervision.

Finally, we would like to implement and evaluate representative-based sampling that groups unlabelled instances within classifier's margin and compute the entropy of each cluster. The unlabelled instance from centroid of the cluster with the highest entropy score (i.e., the densest region) could be considered to be most representative and most uncertain and therefore be selected for wetlab validation. This approach may be more effective than the other active learning approaches evaluated here since it considers the underlying distribution of the unlabelled data.

References

Aha, D. W., Kibler, D. & Albert, M. (1991). Instance-based learning algorithms, *Machine Learning,* 6, 37-66.

Alders,M., Hogan, B.M., Gjini, E., Salehi, F., Al-Gazali,L., Hennekam,E.A., Holmberg,E.E., Mannens,M.M., Mulder,M.F.,Offerhaus,G.J., Prescott,T.E., Schroor,E.J., Verheij,J.B., Witte,M., Zwijnenburg,P.J., Vikkula,M., Schulte-Merker,S. & Hennekam,R.C. (2009). Mutations in CCBE1 cause generalized lymph vessel dysplasia in humans. *Nature Genetics* 41 (12): 1272-1274.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17): 3389-402.

Baram, Y., El-Yaniv, R., & Luz, K. (2004). Online Choice of Active Learning Algorithms. *Journal of Machine Learning Research 5,* 255-291.

Basu, S., Plewczynski, D. (2010). AMS 3.0: prediction of post-translational modifications. *BMC Bioinformatics*, 11:210.

Barton, C. A., Gloss, B.S., Qu, W., Statham, A. L., Hacker, N. F., Sutherland, R.L, Clark, S. J., & O'Brien., P. M. (2010). Collagen and calcium-binding EGF domains 1 is frequently inactivated in ovarian cancer by aberrant promoter hypermethylation and modulates cell migration and survival. *British Journal of Cancer*, 102(1): 87–96.

Bennett, K. P., & Demiriz, A. (1998). Semi-Supervised Support Vector Machines. Advances in Neural Information Processing systems, 12, M.S. Kearns, S.A. Solla, D.A. Cohn, editors, MIT Press, Cambridge, MA, 368 – 374.

Bergamaschi, D., Samuels, Y., O'Neil, N.J., Trigiante, G., Crook,T., Hsieh, J.K., O'Connor, D.J., Zhong, S., Campargue,I., Tomlinson, M.L., Kuwabara, P.E., & Lu,X. (2003). iASPP oncoprotein is a key inhibitor of p53 conserved from worm to human, *Journal of Nature Genetics 33*,162–167.

Bishop, C.M. (1995). Neural networks for pattern recognition. Clarendon Press, Oxford,

Burges, C.J.C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery,* 2(2):121-167.

Byvatov, E. (2011). Introduction to the theory of the Support Vector Machine. http://gecco.org.chemie.uni-frankfurt.de/svmwebc/technical.html (Accessed September 5, 2011).

Campbell, C., Cristianini, N. & Smola, A. J. (2000). Query Learning with Large Margin Classifier. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning,* 111-118.

Chapelle, O. & Vapnik, V. (2000). Model Selection for support vector machines. *Advances in Neural Information processing Systems,* 230-236.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2(3): 1-27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cockman, M. E., Webb, J. D., Kramer, H.B., Kessler, B.M., & Ratcliffe, P.J. (2009a). Proteomics-based Identification of Novel Factor Inhibiting Hypoxia-inducible Factor (FIH) Substrates Indicates Widespread Asparaginyl Hydroxylation of Ankyrin Repeat Domain-containing Proteins. *Journal of Molecular and Cellular Proteomics.* 8(3): 535–546.

Cockman, M.E., Webb, J.D., Ratcliffe, P.J. (2009b). FIH-dependent asparaginyl hydroxylation of ankyrin repeat domain-containing proteins. *Annals of the New York Academy of Sciences*,1177:9-18.

Cortes, C. & Mohri, M. (2004). AUC Optimization vs. error rate minimization. *In Advances in Neural Information Processing Systems*.16(16):313 – 320.

Cortes, C. & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning.* 20(3):273-297.

Cristianini, N. & Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge, Cambridge University Press.

Cui, B., Lin, H. & Yang, Z. (2009). Uncertainty Sampling-based active learning for protein-protein Interaction extraction from biomedical literature. *Expert Systems with Applications*. 36(7):10344 – 10350.

Doyle, S., Monaco, J. & Feldman, M. (2009). A Class Balanced Active Learning Scheme that Accounts for Minority Class Problems: Applications to Histopathology. *MICCAI Workshop on Optical Tissue Image Analysis in Microscopy, Histopathology and Endoscopy*, London, UK, 19.

Druck, T., Gu, Y., Prabhala, G., Cannizzaro, L.A., Park, S.H., Huebner, K. & Keen, J.H. (1996). "Chromosome localization of human genes for clathrin adaptor polypeptides AP2 beta and AP50 and the clathrin-binding protein, VCP". *Genomics* 30 (1): 94–95.

Duan, S. & Bahu, S. (2008). Guided Problem Diagnosis through Active Learning. Guided problem diagnosis through active learning. In *Proc. of 5th IEEE International Conference on Autonomic Computing (ICAC)*, Chicago, IL, 45- 54.

Duda, R.O., Hart, P.E. & Stork, D.G. (2001). Pattern Classification. 2nd Edition. Wiley-Interscience.

Elkins, J.M., Hewitson, K.S., McNeil, L.A., Seibel, J. F., Schlemminger, I., Pugh, C.W., Ratcliffe, P.J. & Schofield, C.J. (2003). Structure of factor-inhibiting hypoxia-inducible factor (HIF) reveals mechanism of oxidative modification of HIF-1 alpha. *Journal of Biological Chemistry.* 278(3):1802-1806.

Ertekin, S., Huang, J., Bottou, L. & Giles, L. (2007). Learning on the Border: Active Learning in Imbalanced Data Classification. In: *CIKM'07: Proceedings of the 16th ACM conference on Conference on information and knowledge management,* Lisboa, Portugal, 127 – 136.

Forman, G. (2002). Incremental Machine Learning to Reduce Biochemistry Lab Costs in the Search for Drug Discovery. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD'02), Edmonton, Alberta, 33-36.

Friedman, N., Geiger, D. & Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine Learning,* 29, 131-163.

Gorina, S. & Pavletich, N.P. (1996). Structure of the p53 tumor suppressor bound to the ankyrin and SH3 domains of 53BP2. *Journal of Science*, 274: 1001-1005.

Green, J. R., Korenberg, M. J. & Aboul-Magd, M. O. (2009). PCI-SS: MISO dynamic nonlinear protein secondary structure prediction. *BMC Bioinformatics*, 10:222.

Guo, Y. & Greiner, R. (2007). Optimistic Active-Learning Using Mutual Information. In *Proceedings of International Joint Conference on Artificial Intelligence*, Springer, 823 – 829.

Guoliang, L. (2009). Knowledge discovery with Bayesian Networks. PhD Thesis, National University of Singapore. http://tinyurl.com/3lwm7zx, (Accessed July 25, 2011).

Hand, D. J. & Yu, K. (2001). Idiot's Bayes:Not So Stupid after All?. *International Statistical Review* 69(3):385-398.

Hardy, A.P., Prokes, I., Kelly, L., Campbell, I.D. & Schofield, C.J. (2009). Asparaginyl beta-hydroxylation of proteins containing ankyrin repeat domains influences their stability and function. *Journal of Molecular Biology,* 392(4):994-1006.

Hennekam, R. C. M., Geerdink, R. A., Hamel, B. C. J., Hennekam, F. A. M., Kraus, P., Rammeloo, J. A. & Tillemans, A. A. W. Autosomal recessive intestinal lymphangiectasia and lymphedema, with facial anomalies and mental retardation. *American Journal of Medical Genetics*. 34: 593-600.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2010). A practical Guide to Support Vectors Classification.http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf. (Assessed August 05, 2011).

Hu, R., Mac Namee, B. & Delany, S.J. (2010). Off to a good start: Using clustering to select the initial training set in active learning. In: *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference*, Daytona Beach, Florida, 26 -31.

Huang, E.L. & Bunn, F.H. ((2003). Hypoxia-inducible Factor and Its Biomedical Relevance. *Journal of Biological Chemistry*. 278(22):19575-19578.

Huang, J. & Ling, C.X. (2005). Using AUC and Accuracy in evaluating Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering-TKDE*, 17(3):299-310.

Quinlan, J.R. (1993). C45: programs for machine learning, Morgan Kaufmann, San Mateo, California.

Jiang, J., & Horace H S Ip. (2008). Active Learning for the Prediction of Phosphorylation Sites. *Proceedings of International Joint Conference on Neural Networks (IJCNN2008) and 2008 IEEE World Congress on Computational Intelligence (WCCI2008),* Hong Kong, 3158-3165.

Kang, J., Ryu, R.K. & Kwon, H –C. (2004). Using Cluster-Based Sampling to Select Initial Training set for Active Learning in Text Classification. *The 8th Pacific-Asia Conference on Knowledge Discovery and Dat Mining*, Springer-Verlag Berlin Heildelberg, 384-388.

Khan, D., M., & Mohamudally, N. (2010). An Agent Oriented Approach for Implementation of the Range Method of Initial Centroids in the K-Means clustering Data Mining Algorithm. *International Journal of the Information Processing and Management*, 1(1):104-113.

King, R.D., Whelan, K.E., Jones, F.M., Reiser, C.H., Bryant, C.H., Muggleton, S.H., Kell, D.B. & Oliver, S.G. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. *Journal of Nature,* 427(6971):247-52.

Kittler, J.T., Chen, G., Kukhtina, V., Vahedi-Faridi, A., Gu, Z., Tretter, V., Smith, K.R., McAinsh, K., Arancibia-Carcamo, I.L., Saenger, W., Haucke, V., Yan, Z. & Moss, S.J. (2008). Regulation of synaptic inhibition by phospho-dependent binding of the AP2 complex to a YECL motif in the GABAA receptor gamma2 subunit. *Proceedings of the National Academy of Sciences of the United States of America*, 105(9): 3616-3621.

Lancaster, D.E., McDonough, M.A., Schofield, C.J. (2004). Factor inhibiting hypoxia-inducible factor (FIH) and other asparaginyl hydroxylases. *Biochemical Society Transactions*, 32(Pt6):943-945.

Lando, D., Peet, D. J., Whelan, D.A., Gorman, J. J. & Whitelaw, M.L. (2002). Asparagine hydroxylation of the HIF transactivation domain a hypoxic switch. *Journal of Science*, 295(5556):858-61.

Laska, M.J., Vogel, U., B., Jensen, B.U., Nexo, A. B. (2010). P53 and PPP1R13L (alias iASPP or RAI) form a feedback loop to regulate genotoxic stress responses. *Biochimica et Biophysica Acta*, 1800:1231-1240.

Lee, T.-Y., Huang, H. –D., Hung, J. –H., Huang, H. –Y., Yang, Y. –S. &Wang, T. –H. (2006). dbPTM: an information repository of protein post-translational modification. *Journal of Nucleic Acids Research,* 34:622-627.

Lewis, D., Gale, W. (1994). A sequential Algorithm for Training Text Classification. *Proceedings of the Eleventh International Conference on Machine Learning,* New Brunswick, NJ, 148-156.

Liu, Z. (2009). Computational *Prediction* of Asparagine and Aspartate *Hydroxylation* Sites on Human Proteins. M.Sc. Thesis, *Carleton University*, Ottawa, Canada.

Liu, Y. (2004). Active Learning with support vectors machine Applied to Gene Expression Data for Cancer Classification. *Journal of chemical information and computer sciences*, 44(6):1936-1941.

Lodish, H. F., Berk, A., Kaiser, C. A., Krieger, M., Scott, M. P., Bretscher, A., Ploegh, H. & Matsudaira, P. (2008). Molecular Cell Biology, 6th ed., W.H. Freeman and Company, New York, USA.

Lowe, S.W., Lin, A.W. (2000). Apoptosis in cancer, *Carcinogenesis* 21:485–495.

Luo, T., Kramer, K., Goldgof, D.B., Hall, L.O., Samson, S.,Remsen, A. & Hopkins, T. (2005). Active Learning to Recognize Multiple Types of Plankton. *Journal of Machine Learning Research,* 6, 589-613.

Martinez-Ramon, M. & Christodoulou, C. (2006). Support Vector Machines for Antenna Array Processing and Electromagnetics. Morgan & Claypool Publishers, CO, USA.

Mohamed, T.P., Carbonell, J.G. & Ganapathiraju, M.K. 2010. Active learning for human protein-protein interaction prediction. *BMC Bioinformatics*, II (supl, 1):S57.

Nakagawa, H., Koyama, K., Murata, Y., Morito, M., Akiyama, T. & Nakamura, Y. (2000). "APCL, a central nervous system-specific homologue of adenomatous polyposis coli tumor suppressor, binds to p53-binding protein 2 and translocates it to the perinucleus". *Cancer Research* 60 (1): 101–5.

Naumovski, L., Cleary, M.L. (1996). The p53-binding protein 53BP2 also interacts with Bc12 and impedes cell cycle progression at G2/M. *Molecular Cell Biology*, 16(7):3884-92.

Nguyen, H., Smeulders, A. (2004). Active learning using preclustering. In *Proc. 21st International Conference on Machine Learning*, Banff, Canada. 2004, 623-630.

Northwest Association for Biomedical Research. Wet Lab, DNA Barcoding: From Samples to Sequences.http://nwabr.org/sites/default/files/learn/bioinformatics/AdvWetLab.pdf (Accessed August 20, 2011).

Olsson, F., Tomanek, K. (2009). An Instrinsic Stopping Criterion for Committee-Based Active learning. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL),*Buolder, Colorado, 138-146.

Tomanek, K. (2010). Resource-Aware Annotation through Active Learning. PhD Thesis, Technical University of Dortmund. 2010. https://eldorado.tu-dortmund.de/bitstream/2003/27172/1/dissertation_katrin_tomanek_final_color.pdf (Accessed July 25, 2011).

Peet, D.J. & Linke, S. Regulation of HIF:asparaginyl hydroxylation. (2006) In Signalling Pathways in Acute Oxygen Sensing, Novartis Foundation Symposium No. 272, D. J. Chadwick and J.Goode, ed. (Weinheim, Germany: Wiley-VCH Verlag GmbH).

Peet, D. J., Lando, D.J., Whelan, D.A., Whitelaw, M. L & Gorman, J. J. (2004). Oxygen-dependent asparagine hydroxylation. *Methods Enzymology*, 381,467-87.

Promega Protocols and Applications Guide. www.promega.com (Accessed August 16, 2011).

Quinlan, J.R. (1993). C4.5: Program for machine learning, Morgan Kaufmann, San Mateo, California.

Ratsch, G. & Warmuth, M.K. (2003). Efficient Margin Maximizing with Boosting. *Journal of Machine Learning Research,* 6, 2131 – 2152.

Robinson, R.A., Lu, X., Jones, E.Y. & Siebold, C. (2008). Biochemical and structural studies of ASPP proteins reveal differential binding to p53, p63, and p73. *Structures,* 16(2):259-68.

Rodriguez,J., Gupta, N., Smith, R. & Pevzner, P. (2008). Does trypsin cut before proline? *Journal of Proteome Research,* 7, 300-305.

Rosset, S., Zhu, J. & Hastie, T. (2003). Margin Maximizing Loss Functions. In *Proceedings of the 17th Annual Conference on Neural Information Processing Systems (NIPS),* Whistler, B.C., 1237 – 1246.

Roy, N. & McCallum, A. (2001). Toward optimal active learning through sampling estimation of error reduction. *Proceedings of the 18th International Conference on Machine Learning,* Morgan Kaufmann, San Francisco , CA, 441–448.

Samuels-Lev, Y., O'Connor, D.J., Bergamaschi, D., Trigiante, G., Hsieh, J.K., Zhong, S., Campargue, I., Naumovski, L., Crook, T. & Lu, X. (2001). "ASPP proteins specifically stimulate the apoptotic function of p53". *Molecular Cell Biology* 8 (4): 781–94.

Schohn, G. & Cohn, D. (2000). Less is More: Active Learning with support Vector Machines. *Proceedings 17th International Conference on Machine Learning.* San Francisco, CA, 839 – 846.

Settles, B. (2010). Active Learning Literature Survey, Computer Sciences Technical Report 1648, University of Wisconsin-Madison, http://www.cs.cmu.edu/~bsettles/pub/settles.activelearning.pdf (Accessed July 21,2011).

Seung, H.S., Opper, M. & Sompolinsky, H. (1992). Query by committee. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, Pittsburgh, PA, 287–294.

Shen, D., Zhang, J., Su, J., Zhou, G., Tan, C.-L. (2004). Multi-criteria-based Active Learning for Named Entity Recognition. In *proceedings of the 42nd Annual Meeting Association Computing Linguistics*, Barcelona, Spain, 589-596.

Shigeo, A. (2005). Support Vector Machines for pattern Classification. Springer-Verlag London Limited.

Singh, M. (2008). First Year PhD Transfer Report: Active Learning for Image Analysis.http://mlg.ucd.ie/~smohan/downloads/PhDTransferReport08.pdf (Accessed June 30, 2011).

Tax, D., de Ridder, D. & Duin, R. P. W. (1997). Support vector classifiers: a first look. *Proceedings of the 3rd Annual Conference on the Advanced School for Computing and Imaging (ASCI'97)*, Heijen, Netherland, 253-258.

Tidow, H., Andreeva, A., Rutherford, T.J. & Fersht, A.R. (2007). "Solution structure of ASPP2 N-terminal domain (N-ASPP2) reveals a ubiquitin-like fold". *Journal of Molecular Biology.* 371 (4): 948–58.

Tong, S. & Koller, D. 2001. Support vector Machine active learning with application to text classification. *Journal of Machine Learning Research*, 2, 45-66.

Vapnik, V.N. (1999).The nature of statistical learning theory (2nd edition), Springer, New York.

Vlachos, A. (2008). A Stopping Criterion for Active Learning. *Computer Speech and Language*. 22(3):295-312.

Voet, D., Voet, J. G. & Pratt, C. W. (2009). Fundamentals of Biochemistry: Life at the Molecular level, John Wiley & Son, Hoboken, NJ.

Warmuth, M.K., Ratsch, G., Mathiesib, M., Liao, J. & Lemmen, C. (2002). Active Learning in the Drug Discovery Process. *Advances in Neural Information Processing Systems*, 2, 1449-1456.

Xu, H., Wang, X., Liao, Y. & Zheng, C. (2009). An Uncertainty sampling-based Active Learning Approach for Support Vector Machines. *International Conference on Artificial Intelligence and Computational Intelligence,* Shanghai, China, 208-213.

Xu, Z., Yu K., Tresp, V., Xu, X. & Wang, J. (2003). Representative of sampling for text classification using SVM. In *Proceedings of the Advances in Information retrieval, 25th European conference on IR research (ECIR),* Italy, 2633, 393-407.

Yuan, W., Han, Y., Guan, D., Lee, S. & Lee, Y. (2011). Initial Training Data Selection for Active Learning. *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication,(ICUIMC 2011*), Seoul, Korea, Feb 21 - 23, 2011.

Ze-Jun, L., Lu, X. & Zhong, S. (2005). ASPP—Apoptotic specific regulator of p53. *Biochimica et Biophysio Acta (BBA)-Reviews on Cancer,* 1756 (1):77-80.

Zhang, N., Fu, Z., Linke, S., Chicher, J., Gorman, J.J., Visk, D., Haddad, G.G., Poellinger, L., Peet, D.J., Powell,F. & Johnson, R.S. (2010). The Asparaginyl Hydroxylase Factor Inhibiting HIF-1α Is an Essential Regulator of Metabolism. *Cell Metabolism* 11(5):364-378.

Zhu, J. Wang, H., Tsuo, B.K. & Ma, M. (2008). Active Learning with Sampling by Uncertainty and Density for Data Annotations. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1323-1331.

Zhu, J. & Hovy, E. (2007). Active Learning for word Sense Disambiguation with Methods for Addressing the Class Imbalance Problem. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Proceeding and Computational Natural Language Learning*, Prague, 783-790.

Zhu, J., Wang, H. & Hovy, E. (2008). Learning a Stopping Criterion for Active Learning for Word Sense Disambiguation and Text Classification.

Invitrogen. Achieve 99% transfection Efficiency in a single stroke. *http://toolszh.invitrogen.com/content/sfs/brochures/710_011031_LipoFect_bro.pdf* (Accessed August 23, 2011).

OpenBiosytems RNAi, Gene Expression, Antibodies. (Accessed September 13, 2011).

APPENDIX A: GENETIC CODE AND PROTEIN SEQUENCE

A.1 Genetic Code

The genetic code is the set of rules by which information encoded in genetic material (DNA or mRNA sequences) is translated into proteins (amino acid sequences). The codon AUG for methionine serves as an initiation site, i.e. the first AUG in an mRNA's coding region is where translation into protein begins. A = adenine, G = guanine, C = cytosine, T = thymine, and U = uracil.

nonpolar polar basic acidic (stop codon)

1st base	2nd base U		2nd base C		2nd base A		2nd base G	
U	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine
	UUC	(Phe/F) Phenylalanine	UCC	(Ser/S) Serine	UAC	(Tyr/Y) Tyrosine	UGC	(Cys/C) Cysteine
	UUA	(Leu/L) Leucine	UCA	(Ser/S) Serine	UAA	Stop (*Ochre*)	UGA	Stop (*Opal*)
	UUG	(Leu/L) Leucine	UCG	(Ser/S) Serine	UAG	Stop (*Amber*)	UGG	(Trp/W) Tryptophan
C	CUU	(Leu/L) Leucine	CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine
	CUC	(Leu/L) Leucine	CCC	(Pro/P) Proline	CAC	(His/H) Histidine	CGC	(Arg/R) Arginine
	CUA	(Leu/L) Leucine	CCA	(Pro/P) Proline	CAA	(Gln/Q) Glutamine	CGA	(Arg/R) Arginine
	CUG	(Leu/L) Leucine	CCG	(Pro/P) Proline	CAG	(Gln/Q) Glutamine	CGG	(Arg/R) Arginine
A	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine
	AUC	(Ile/I) Isoleucine	ACC	(Thr/T) Threonine	AAC	(Asn/N) Asparagine	AGC	(Ser/S) Serine
	AUA	(Ile/I) Isoleucine	ACA	(Thr/T) Threonine	AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine
	AUG[A]	(Met/M) Methionine	ACG	(Thr/T) Threonine	AAG	(Lys/K) Lysine	AGG	(Arg/R) Arginine
G	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine
	GUC	(Val/V) Valine	GCC	(Ala/A) Alanine	GAC	(Asp/D) Aspartic acid	GGC	(Gly/G) Glycine
	GUA	(Val/V) Valine	GCA	(Ala/A) Alanine	GAA	(Glu/E) Glutamic acid	GGA	(Gly/G) Glycine
	GUG	(Val/V) Valine	GCG	(Ala/A) Alanine	GAG	(Glu/E) Glutamic acid	GGG	(Gly/G) Glycine

Figure A. A.1. The genetic code table, adapted from Wikipedia.

A.2 Sequence Representation of Protein Synthesis

The biological information necessary for labelling strands is the location of the 5'-phosphate group and 3'-hydroxyl group as these ends determine the direction of transcription and translation, i.e. translation proceeds in the 5' to 3' direction.

5'... T A T A G C G T T C A T ... 3'	{DNA template (noncoding) strand, used as a template for transcription}
3'... A T A T C G C A A G T A ... 5'	{DNA nontemplate (coding) strand. Complementary to the template strand}
3'... A U A U C G C A A G U A ... 5'	{RNA strand transcribed from DNA template, identical to DNA nontemplate strand, except all thymines now uracils (T→ U), mRNA}
5'... A U G A A C G C U A U A ... 3'	{exactly the same sequence as above, except, the AUG start codon, mRNA}
↓ Translational of mRNA	
Met – Asn – Ala- Ile -	**Peptide**

Transcription (DNA) ⟶ Translation (RNA) ⟶ Proteins

A.3 Plasmid DNA Transfection efficiency using Lipofectamine 2000 Reagent

Transfection efficiency using Lipofectamine is dependent on various factors such as plasmid DNA, cell line, cell confluency/viability, growth medium and relative surface area. This relative surface area is determined by counting cell relative to the cell well plates. We have used 6-cm

well plate and scaled up to 10-cm well plate for optimized transfection efficiency to achieve maximum protein expression (yield) based on the given Table A.2 (Invitrogen, 2005).

Table A.1. Optimized transfection working Volume

Culture vessel	Surface area per well (cm^2)	Relative Surface Area (vs. 24-well)	Volume of plating medium	DNA (µg) and Dilution Volume (µl)	Lipofectamine™ 2000 (µl) and Dilution Volume (µl)
96-well	0.3	0.2	100 µl	0.2 µg in 25 µl	0.5 µl in 25 µl
24-well	2	1	500 µl	0.8 µg in 50 µl	2.0 µl in 50 µl
12-well	4	2	1 ml	1.6 µg in 100 µl	4.0 µl in 100 µl
35-mm	10	5	2 ml	4.0 µg in 250 µl	10 µl in 250 µl
6-well	10	5	2 ml	4.0 µg in 250 µl	10 µl in 250 µl
60-mm	20	10	5 ml	8.0 µg in 0.5 µl	20 µl in 0.5 µl
10-cm	60	30	15 ml	24 µg in 1.5 µl	µl in 1.5 µl

A.4 Detailed Western blot protocol

A.4.1 Buffer Preparation

Components	12% Resolving Volume (µl)	4% Stacking Volume (µl)
Protogel	3200	520
Resolving Buffer	2080	0
Stacking Buffer	0	960
MilliQ Water	2632	2440
30% APS	28	8
TEMED	8	4

Figure A.1. Composition of SDS-PAGE for 10% resolving and 4% stacking gel for 1.5mm spacer plate

Laemmli 2X buffer

62.5 mM Tris-pH 6.8

2% SDS (Sodium Dodecyl Sulfate)

25% glycerol

0.01% bromophenol blue

50µl β-mercaptoethanol

Tris-Buffered Saline Tween-20 (TBST)

20 mM Tris-HCl, pH 7.6
137 mM NaCl
0.1% Tween 20

Protein transfer buffer (PTB)

20mM Tris-HCL, pH 8.0
150mM Gycine
20% Methanol

Blocking Solution

20 ml TBST buffer
1g carnation non-fat dry milk

Renaissance Western Blot Chemiluminescent substrate/reagent (NEN Life Science Products, Boston, MA)

0.5ml of enhanced luminal reagent
0.5ml oxidizing agent

A.4.2 SDS-PAGE

- Prepare SDS-PAGE gel (Table A1)
- Place gel in SDS-PAGE running apparatus and fill with transfer buffer (PTB)

- Mix 1:1 with 50μL β-mercaptoethanol in 950 μL of Laemmli buffer and add to protein sample
- Heat cell lysate protein at 100°C for 3 minutes
- Load 2 – 5 μL of protein ladder i.e. molecular weight marker
- Run at 150V constant until the blue dye reaches the bottom of the gel electrophoresis apparatus.

A.4.3 Western blotting (protein transfer)

- Cut pieces of Whatman paper slightly smaller than the gel (blotting paper) and Immobilon-P transfer membrane (Millipore, Bedford, MA) of type PVDF.
- Wet membrane in 100% methanol and subsequently submerged the membrane completely in MilliQ water.
- Then, the gel and membrane are sandwiched between sponge and paper in the Western blot cassette as follows: sponge/paper/gel/membrane/paper/sponge, i.e., 2 x blotting sponges, 3 X whatmam paper, SDS-PAGE gel, PVDF membrane, 3 X whatman paper, and 2 X blotting sponges. This sandwiched arrangement must be prepared in 1X Tri-glycine buffer solution also used for running the buffer.
- Transfer Western blotting cassette into Western blot running apparatus and place bio-ice cooling unit with ice in the opposite side of the running apparatus, then, cover lid and connects black to cathode and red cables to the anode.
- Fill Western blot running apparatus with PTB to the top and allows to overflow to the 2 gel mark on the running apparatus.
- Run overnight at 180mA at 4°C

A.4.4 Blocking, antibodies, washing, and incubation

- Incubate membrane in blocking solution for 1 hr at room temperature.
- Dilute the primary antibodies (anti-AP2M1 and anti-TP53BP2) in 1:1,000 dilutions and IgR horseradish peroxidase (HRP)-conjugate secondary antibody (goat anti-rabbit) dilution 1:2,000) in TBST.

- Incubate membrane with primary antibodies for 1hr, remove and store primary antibodies in a 4°C, then wash blot with TBST with mild shaking, 4 X 30 minutes.
- Incubate membrane with secondary antibody for 1hr, remove and store secondary antibody in a 4°C, then wash blot with TBST with mild shaking, 4 X 30 minutes at room temperature.

A.4.5 Protein bands detection or development method

- Drain the remaining wash by dabbing the edge of the blot on a kimwipe until all is gone
- Place membrane in a Saran wrap
- Pipette 1mL of the chemiluminescent substrates into the blot and incubate for one minute.
- Cover blot with the Saran wrap and place in an autorad cassette
- (Dark room) Expose membrane to film for 30 seconds, or longer as required to achieve best protein band signals on the film.
- (Dark room) Develop film for 5 minutes in developing solution, wash film in water for 1 minute and fix for 5 minutes in fixing solution.

A.5 Coomassie stain/destain protocol

A.5.1 Fix-gel and destain Composition

10 % (v/v) acetic acid (10mL x 2)

25% (v/v) methanol (25mL x 2)

Adjust the total volume to 200mL with MilliQ water, i.e., final concentrations are 10mL methanol in water with 20mL acetic acid

A.5.2 Coomassie blue stain Composition

0.25% (w/v) Coomassie blue dye R-250 (Biorad, Hercules, California, USA)

7.5 % (v/v) acetic acid (7.5mL)

50% (v/v) methanol (50mL)

Add 50mL methanol in MilliQ water, i.e., adjust the final concentration to 100mL with 50mL methanol, 7.5mL acetic acid and 0.25% Coomassie blue dye R-250.

A.5.3 Coomassie Blue dye stain/destain procedure

- Transfer electrophoresis gel onto 100mL of the fix-gel solution in approximately 25cm X 40 cm plastic dishes and continue to fix the protein in the gel by incubating for 1 hour at room temperature with gentle shaking. After the incubation period, discard the solution.
- Gently cover the gel with 100mL of the Coomassie stain solution and continue to stain the gel for 2 hours at room temperature. Remove and keep the Coomassie solution for future test.
- Cover the gel with another 100mL of the destain solution, and continue to destain the gel until the protein bands become clear without background staining of the gel.
- Cut required bands of the protein of interest and store gel in Eppendorf tube in -80°C for mass spectrometry analysis.

A.6 Trypsin_digest_script

```
import sys, os,string,re
'''
The script produces a tryptic digestion of a given set of protein sequences. The input file contains
only one letter code amino acid of the considered protein sequence. The algorithm implements
accepted "Keil rule" is that trypsin (enzyme) cleaves next to or after lysine (K) or arginine(R),
unless they are followed by a P. Note: input file must be in FASTA format of the protein
sequence saved as .txt
'''
# Open the file and read it line by line.

myprotein = open(raw_input('Enter input filename: '),'r')
my_protein = []

#check if the output file already exist
if  os.path.exists("trypsin_digest.txt"):
    os.remove("trypsin_digest.txt")
outfile = open("trypsin_digest.txt",'w+')

for protein in myprotein:
     myprotein = protein.rstrip('\n')
     my_protein.append(myprotein)
```

```
    my_pro = (''.join(my_protein))

#Keil rule, trypsin cleaves next to lysine(K) or
#arginine(R),unless (K or R) is followed by a P.
peptides = re.sub(r'(?<=[RK])(?=[^P])','\n', my_pro)
outfile.write(peptides)
print 'results written to:\n', os.getcwd() +'\
trypsin_digest.txt'
```