Classification of Protein Ubiquitination Sites using Hidden Markov Models

Project Proposal, March 20th 2017

Group: 11

Jinny Lee — MASc Student under the supervision of Dr. Andy Adler, Dr. Eran Ukwatta

George Hanna — 4th year Biomedical and Electrical Engineering Student.

Understanding of the Problem

Protein Dataset

- Class Imbalance: positive to negative ratio of 0.164
- Large number of features
- Small subset of labelled data points
- Limited of 1000 new labelled training points
- Lack of background information

Project Plan

Selected Approach: Hidden Markov Models, Combined with 2 other methods

Working Environment: MATLAB 2017a

- **Data Pre-processing**: Scripting
- **Feature Extraction/selection:** Literature derived open source tools
- **Experiment Design:** multiple different data splits and training schemes
- Training Process
- Testing
- **Meta-Learning approaches:** combining HMMs and/or other methods

Current Progress

Literature Review of Current Methods

- Comparison of 5 journal publications on protein ubiquitination.
- Metrics include accuracy, precision, recall.
- Examination of feature selection steps, classifiers used, used of meta-learning.

Exploratory Classification

- Used WEKA to explore 7 classifiers performance on class imbalanced and balanced data (e.g. random forests, naïve Bayes, KNN, HMM)

Current Progress

Environment Setup

- Familiarization with HMM toolbox in MATLAB
- Research into other environments for running HMMs (e.g. GeneMark)
- Exploration of other tools for data-processing
- Brainstorming ideas for data splitting and experiment design