CPSC 290 Project Proposal: Application of Machine Learning to RNA Expression Data

Undergraduate researcher: Ian Gonzalez
Faculty advisor: Mark Gerstein
Postdoc supervisor: Robert Kitchen

Background:

Micro RNAs (miRNAs) are small (19-23nt) RNA molecules that play key regulatory roles in a variety of biological pathways. The functionality of miRNAs has been extensively characterized within certain cell and tissue types – miRNAs generally function by inhibiting the translation of mRNA transcripts or by inducing the destruction of mRNA transcripts. Two of the most common biochemical methods used to characterize levels of expression within a sample are RNA-seq and qPCR. RNA-seq takes advantage of the high sequencing coverage of next-generation sequencing techniques to sequence all the RNA present in a cell at a given time; the qPCR method takes advantage of the standard polymerase chain reaction technique to amplify RNA sequences in the cell. The major difference between these techniques is that RNA-seq provides sequences and relative abundances of all small RNAs, while qPCR gives absolute abundance (but only for one or a few RNAs at a time).

Some recent studies have detected and analyzed miRNAs associated with extracellular spaces (Williams et al. 2013). The function of these extracellular miRNAs (abbreviated as exRNA in this proposal) is largely unknown – some have posited that they may be involved in intercellular signaling, but it is also possible that they serve no substantive purpose. Even if they serve no discernible purpose, exRNAs may be clinically useful for diagnosis – abnormal levels of circulating miRNAs could be indicative of some disease or other ailment. One of the properties that may give some hint at the answer to this question is the tissue (or tissues) of origin of these exRNAs. However, determining this property is no straightforward task – extracellular miRNAs are almost certainly produced by a variety of tissues, and there is no accepted standard for assigning exRNA expressions (or relative expressions) to specific cell types. It is possible that existing tissue-specific cellular miRNA expression data from RNA-seq and qPCR experiments contain enough information to define miRNA signatures for each tissue type and may allow us to make predictions about exRNA origins. We propose that exRNA
expression profiles may be able to be ‘decomposed’ by a mixture of cellular miRNA signatures. If this is the case, it may also be possible to infer the relative contributions of each tissue type to the exRNA pool.

**Project description:**

The Gerstein lab is currently investigating a variety of computational data analysis and machine learning techniques to search for the traces of tissue-specific RNA expression within the exRNA samples. Ideally, the analysis will produce a definitive set of miRNA expression “fingerprints” that differentiate the tissues that have been studied by researchers so far – these will likely end up being sets of relative miRNA expression levels that are seen quite often in certain tissues and not as often in other tissues. Assuming that sufficient tissue-discriminatory information exists in the data set, these miRNA “fingerprints” will then be used for training a classifier that, when given a set of miRNA expressions (such as the exRNA data), will predict which tissue the miRNAs originated in.

I have already undertaken some initial analysis of public RNA-seq miRNA expression data sets and the results are promising. Data from 14 different human tissues were collected from the literature and processed with a smallRNA analysis pipeline developed in the Gerstein Lab. After feature scaling, a hierarchical clustering algorithm was used to find 11 highly correlated miRNA clusters within the data – a simple analysis of variance revealed that the relative expression levels within these clusters also correlated quite well with tissue type. The “fingerprints” of each tissue have yet to be determined, but a simple visual analysis of the results suggests that multi-miRNA fingerprints will likely be highly discriminative.

The next step is to perform a similar analysis of existing qPCR expression data. These results will be compared with the RNA-seq results for the purposes of validation and will eventually be used to train a qPCR-specific classifier that will be bundled with the RNA-seq classifier in the final product. The qPCR miRNA data will be taken from the miRNAbodymap project (Mestagh, et al. 2011), which aggregated and normalized tissue-specific expression data for the purposes of improving miRNA function prediction.

With this analysis complete, we will use the results to inform efficient feature selection for the classifiers that will be trained on both data sets. Given the high levels of correlation between some miRNAs, it’s possible that we will include dimensionality reduction as a standard preprocessing step – we will also explore methods for including the most discriminatory miRNA...
ratios as features within the classification system. A variety of machine learning techniques will then be tested on the data to ensure a high level of training and cross-validation accuracy. It is difficult to know exactly what kind of system will be used before our analysis is complete, but neural networks are expected to perform well considering the number of features in the data set if all miRNAs are used.

Finally, the exRNA data sets will be fed into the classifier and the results analyzed for any interesting patterns that may be relevant to the study of exRNA function. After our assessment is complete, the best classifiers for both data sets will be packaged along with any other necessary pipeline steps (mainly preprocessing) in order to create an open source tool that other researchers can use for similar analysis of unknown or mixed samples of miRNAs.

**Discrete Project Goals (Deliverables):**

The goals of this project are as follows:

1. Aggregate tissue-specific data on miRNA expression levels from both RNAseq and qPCR experiments (already done for RNAseq).
2. Use data analysis techniques to find the miRNA expression features that best differentiate various tissues.
3. Taking the features found in part 2 into account, develop a machine learning pipeline that, given a new set of miRNA expression data from a qPCR or RNA-seq experiment, will return a set of probabilities that the miRNAs come from each tissue involved in the classifier’s training. This pipeline will be used to analyze existing exRNA data and will also be released as free open source software for other researchers to use. (This software may come in the form of a web tool, an R package, or both).

**References:**