Identification of LncRNA Characteristics and Associations Using HIN Based Techniques

Abstract

The successful completion of *Human Genome Project* resulted in revolutionary developments in sequencing and other instrumentation technologies in the field of life science. This catalysed the generation of tremendous volumes of data, growing super-exponentially every year. Consequently, the application of information technology to the field of biology emerged as an absolute necessity. This led to the evolution of bioinformatics, a multi-disciplinary domain that integrates the areas of computer science, statistics, biology and so on. Bioinformatics aims to manage and process the massive, heterogeneous and ever-increasing data generated from high-throughput biological experiments. It addresses research issues from biology through the application of reliable computational techniques for fast and better elucidation of living systems.

Long non-coding RNA (LncRNA) has become one of the most sought-after terminologies in current day biological research. Until a decade ago, the non-coding regions of RNA were broadly considered as transcriptional noise with no defined functions. Recent studies refuted the misconceptions and categorically revealed that lncRNAs are actively involved in innumerable biological processes and associated to a wide variety of diseases. This knowledge prompted biological experiments to focus on lncRNAs with an aim to decipher their activities in cellular processes and diseases.

Unlike genes, proteins and other RNAs, lncRNAs have very low sequence conservation and poor structure-function association. These unique properties make the traditional wet-lab study of lncRNAs tedious and expensive. Moreover, the increased interest on lncRNAs led to the identification of newer lncRNAs on a regular basis, whereas, their annotation process to determine the biological functions has not achieved the necessary pace. Under these circumstances, the development of computational methods, tools and algorithms to unfold lncRNA characteristics deserves special attention. Considering these facts, this thesis proposes mainly three machine learning based computational models to uncover biological activities, functions and disease associations of lncRNAs using the principles of Heterogeneous Information Networks (HIN).

The first contribution of this research work is an HIN based model that predicts the cancer cell activity of lncRNAs. The patterns of interaction with proteins and coexpression of lncRNAs are mapped to HIN and a meta-path based feature set is constructed. Using this, an SVM classifier is trained which predicts the lncRNAs as oncogenic or tumor suppressor. Upon k-fold cross validation, the model showed a prediction accuracy of 0.83. Further, the model showed an accuracy of 0.80 for an independent test set of unseen samples. The model was able to predict the cancer cell activity of many lncRNAs which are validated from recent literature.

The second contribution is to predict the functions of lncRNAs using an HIN based model constructed from lncRNA coexpressions, associations with proteins and existing functions as well as protein-protein interactions. The meta-path based parameter *AvgSim* is applied to construct the feature set for an RF classifier which identified novel associations of lncRNAs and functions. The proposed model outperformed many of the state-of-the-art models in terms of statistical parameters such as precision and f-score. The prediction accuracy obtained was 0.74. Many novel associations are validated using experimental results taken from recent literature. A separate case study of two well-known lncRNAs (HOTAIR, H19) is conducted and found that many associations predicted by the model are really existing.

Even though a number of techniques exist to predict lncRNA-disease associations, no computational approach exist in the literature that associates lncRNAs to pathways. To fill this gap an HIN based model is proposed as third contribution which predicts lncRNA-pathway associations. The model first predicts lncRNA-disease associations and then extends it to pathways by enriching the gene set associated to each lncRNA through diseases. The existing lncRNA-disease associations are mapped to an HIN and novel associations are inferred using an SVM classifier. A new parameter, *Association Index* is proposed to construct the feature set for the classifier. The accuracy obtained for lncRNA-disease association prediction was 0.78. The model identified 430 new associations for which experimental evidence existed. Additionally, it revealed many lncRNA-pathway associations, for which proofs were obtained from the litera-

ture. Interestingly, more than one pathways associated with common lncRNAs were found to be interdependent as well. The role of lncRNAs in determining such interactions needs further exploration.

Apart from these three major contributions, an algorithm namely RelRank is also proposed to rank the meta-paths in an HIN based on their relevance. The algorithm assigns a relevance score to each meta-path in the HIN and ranks them based on this score. The correctness of the algorithm is verified by applying it on a real biological HIN.

The computational methods proposed in the thesis take lncRNA research one step further. In addition, the contributions do extend the researches where lncRNA is used as therapeutic target or disease biomarker. The findings here can serve as a stepping stone to even more exciting discoveries that can take us deeper into the fascinating domain of lncRNAs.

Keywords: Long Non-coding RNA, Heterogeneous Information Network, Meta-path, Disease, Pathway, Classification, Support Vector Machine, Random Forest, Cancer