Faculty of Engineering and Information Technology
University of Technology Sydney

Disease Gene Recognition and Editing Optimization Through Knowledge Learned from Domain Feature Spaces

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Hui Peng

May 2019

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This research is supported by the Australian Government Research Training Program.

Signature of Candidate

Production Note:
Signature removed
prior to publication.

Acknowledgments

Foremost, I hope to express my sincere gratitude to my supervisor Prof. Jinyan Li for his continuous support during my PhD study and research, for his patience, motivation, and immense knowledge. His guidance helped me improve both my research skills and other necessary abilities such as scientific writing, academic communicating and presentation skills. The completion of this thesis and related researches would not be possible without his constructive advice for improving them and his valuable time and efforts to make them perfect.

I would like to appreciate Dr. Tao Liu and Prof. Gyorgy Hutvagner, two of my research partners, for their patience to discuss the problems of our research and for their insightful suggestions to finish the research tasks. Many thanks to my co-supervisor Prof. Dacheng Tao for his kind suggestions and help to a part of my research work.

I really appreciate our two team members Yi and Chaowang, who are also my friends and roommates, for their accompany during the past three and a half years, for their great help in not only my study and research but also my life abroad. I am grateful to the three former team members Dr. Jing Ren, Dr. Renhua Song and Dr. Shameek Ghosh, for their warm responses when I requested help from them during the first two years in UTS and even after their graduation. I also thank the other team members Zhixun, Yuansheng, Xiaocai, Xuan and Tao, who joined us in recent two years. I feel very happy to meet all of you in UTS and hold a lot of fantastic activities with you. Those happy moments will be kept in my mind forever.

In addition, I gratefully acknowledge the funding sources, including the ARC Discovery Scholarship and the International Research Scholarship provided by Graduate Research School, that made my PhD work possible. Thanks to the staffs of Advanced Analytics Institute and School of Software, whose work provided so many conveniences to my study and research in UTS.

At last, I really appreciate my father Hongbing Peng, my mother Hejiao Li, my sister Min Peng and My girlfriend Ruqian Peng for their financial and emotional support during my overseas study, for their encouragements when I encountered difficulties and for their concerns to beat my homesickness. My relatives and friends in China, though I have not mentioned you one-by-one, I sincerely thank your frequent greetings and blessings, which make me do not feel lonely. Thanks very much!

Hui Peng
May 2019 @ UTS

Contents

Certificate . . . . **i**

Acknowledgment . . . . **iii**

List of Figures . . . . **xi**

List of Tables . . . . **xix**

List of Publications . . . . **xxi**

Abstract . . . . **xxv**

Chapter 1 Introduction **1**

1.1 Background 1

1.1.1 Protein-coding gene, non-coding gene and non-coding RNA 1

1.1.2 Non-coding RNA and human diseases 3

1.1.3 The CRISPR/Cas9 system 4

1.1.4 CRISPR/Cas9 system design for disease gene editing . 6

1.1.5 Machine learning in bioinformatics 7

1.2 Research Questions and Formulations 9

1.3 Research Contributions 11

1.4 Thesis Structure 15

Chapter 2 Related Work and Literature Review **17**

2.1 Disease-ncRNA Association Prediction 17

2.1.1 Network analysis methods 18

2.1.2 Similarities or semi-supervised methods 22

2.1.3 Supervised learning methods 23

2.1.4 Other types of prediction methods 25
2.2 CRISPR/Cas9 On-target Cutting Efficiency Prediction 26
2.2.1 Binary classification of the sgRNAs 26
2.2.2 Regression methods for sgRNA cutting efficiency prediction 29
2.2.3 A Non-machine learning method for on-target cutting efficiency prediction 31
2.3 CRISPR/Cas9 Off-target Site Detection 31
2.3.1 Wet-lab technologies for off-target site detection 31
2.3.2 Mismatch information scoring methods for off-target site detection 34
2.3.3 Machine learning methods for off-target site detection . 36
2.4 Limitations of Existing Methods 37
2.4.1 Limitations of the disease-ncRNA association prediction methods 37
2.4.2 Limitations of the methods for gene editing optimization 38
2.5 Summary 40

Chapter 3 Precomputed Kernel Matrix SVM Method for Predicting Disease Related miRNAs 41
3.1 Introduction 41
3.2 Method 42
3.2.1 Data sets for the diseases, miRNAs and their related genes 42
3.2.2 Positive samples and negative samples for training the prediction model to identify unknown disease-miRNA associations 43
3.2.3 Precomputed kernel matrices for the support vector machine 45
3.2.4 Measuring the pairwise similarities of diseases or miRNAs 47
3.3 Results 49
3.3.1 The optimal precomputed kernel matrix and the prediction performance 49

3.3.2 Effect of the size of the negative samples on the prediction performance . . . 52
3.3.3 Performance comparison when changing the approach of selecting negative samples . . . 54
3.3.4 Performance comparison: prediction of disease-miRNA relationships by different methods . . . 55
3.3.5 The predicted miRNAs that are related to breast and prostate cancer: Case studies . . . 57
3.4 Conclusion . . . 62

Chapter 4 Cross Disease Analysis of Co-functional microRNA Pairs on A Reconstructed Network of Disease-gene-microRNA Tripartite . . . 63
4.1 Introduction . . . 63
4.2 Method . . . 68
4.2.1 Reconstructing the DGR tripartite network . . . 68
4.2.2 Scoring the multi-disease associated co-functional miRNA pairs . . . 69
4.2.3 Determining the potential co-functional target genes . 71
4.3 Results . . . 71
4.3.1 Multi-disease associated co-functional miRNA pairs and their common dysfunctional target genes . . . 71
4.3.2 An in-depth analysis of five co-functional miRNA pairs 76
4.4 Conclusion . . . 81

Chapter 5 Chromosome Preference of Disease Genes and Vectorization for the Prediction of Non-coding Disease Genes . . . 82
5.1 Introduction . . . 82
5.2 Materials and Methods . . . 87
5.2.1 Diseases, disease genes and KEGG pathways . . . 87
5.2.2 Associations between diseases and lncRNAs . . . 88

5.2.3 Disease gene chromosome preference analysis and disease vectorization method 89
5.2.4 Prioritizing disease related lncRNA genes 93
5.3 Results 95
5.3.1 Chromosome preference and disfavor of disease genes . 95
5.3.2 Performance on the prediction of highly similar diseases using our disease vector representation 98
5.3.3 Performance on the prediction and prioritization of disease related lncRNA genes 101
5.3.4 Performance comparison and case studies 104
5.4 Conclusion 108

Chapter 6 CRISPR/Cas9 Cleavage Efficiency Regression Through Boosting Algorithms and Markov Sequence Profiling 109

6.1 Introduction 109
6.2 Materials and Methods 113
6.2.1 High throughput genome engineering datasets for building the regression and classification models 113
6.2.2 Features for building the regression and classification models 115
6.2.3 Procedures for training our TSAM 119
6.3 Results 121
6.3.1 Nucleotide and cleavage preferences of highly efficient sgRNAs as revealed by the boosting algorithm 121
6.3.2 Further performance improvement by integrating pHMM properties 123
6.3.3 Results on 11 benchmark datasets comparing with the state-of-the-art methods 124
6.3.4 Performance of TSAM on more datasets related to the U6 and T7 expression system 126
6.3.5 Case study: designing sgRNAs for gene therapy 129
6.4 Conclusion 130

Chapter 7 Recognition of CRISPR/Cas9 Off-target Sites Through Ensemble Learning of Uneven Mismatch Distributions 132

7.1 Introduction 132

7.2 Materials and Methods 137

7.2.1 Datasets for training and testing the prediction model . 137

7.2.2 Integrative characteristics of sequence pairs 139

7.2.3 Convert a sequence pair $\langle onTseq, offTSeq \rangle$ into a feature vector 141

7.2.4 Build the prediction model for detection of off-target sites 143

7.3 Results 145

7.3.1 GC count change, 5'-end editing potential and preference145

7.3.2 Off-target site prediction and performance comparison with other methods 147

7.3.3 Comparison of the off-target sites detected by the computational methods and those by the high-throughput sequencing methods 149

7.3.4 Selecting optimal sgRNAs for curing diseases: Two case studies 152

7.4 Conclusion 155

Chapter 8 Conclusions and Future Work 156

8.1 Conclusions 156

8.2 Future Work 159

Chapter A Appendix: Methodology foundation 163

A.1 Adopted Mathematical and Statistical Conceptions 163

A.1.1 Information entropy 163

A.1.2 Fisher's exact test 164

A.1.3 Two-sample Kolmogorov−Smirnov test 164

A.2 Applied Machine Learning Algorithms 165

A.2.1 Support vector machine 165

A.2.2 Ensemble SVM 165

A.2.3 XGBoost 166

A.3 Cross-validation and Performance Indicators 166

A.3.1 Cross-validation 166

A.3.2 Performance indicators 167

Chapter B Visited databases 169

Chapter C Supplementary files 171

Chapter D Appendix: List of Symbols 172

Bibliography 179

List of Figures

1.1 **An example of a CRISPR/Cas9 system cutting a genome DNA sequence.** 6

1.2 **Thesis Structure.** It includes four main parts: Introduction; Related work; My own work; Conclusion and future work. The overview of the contents in each part is shown at the right side. 16

3.1 **Performances of the predictions under different precomputed kernel matrix and α.** We mainly compare the AUC values and the F1 scores of each models with different parameters. K1, K2 and K3 represent the three kernel matrix types such as the average type, the squared root type and the center distance type respectively. The results indicate that the model with the squared root type of kernel matrix and $\alpha = 0.8$ achieves better performance. 51

3.2 **The ROC curves of the permutation test.** The experiment includes the test group and the control group parts. The test group part used the permutated labels for the training samples while the control group part uses the original labels of the same training dataset. Both of the two parts of the experiment adopts our optimal prediction model. 52

3.3 **Performances of the prediction models with different size ratio of negative and positive samples.** The prediction model was trained on the sample sets with different ratio of negative and positive samples. The x-axis shows the ratios. AUC and mcc values were computed based on 10-fold cross validation. The Accuracy is the percentage that the samples in the validation dataset (a dataset with just positive samples but does not overlap with the training sample sets) are predicted correctly. 53

3.4 **The ROC curves of our model compared with RLSMDA based on the same positive samples.** The comparison is based on the same positive sample set and the different prediction model of RLSMDA and our newly designed model. The average AUC value of our model is 0.9896 while the RLSMDA obtains the lower value of 0.9475. 57

3.5 **The top 30 predicted breast cancer-miRNA and prostate cancer-miRNA associations and the verification resources.** The left part shows the predicted breast cancer related miRNAs and the right part gives the predicted prostate cancer related miRNAs. The labels of the edges illustrate the ranks of the predicted associations and the confirming types. The characters “*”, “#” or “$” stand for that the corresponding associations can be confirmed by the records in miR2Disease , HMDD or miRCancer respectively. The character “@” means that the association can be confirmed by other articles. A co-functional pair miR-195-5p-miR-15b-5p is highlighted. . . . 59

3.6 **The percentages of the predicted disease-miRNA associations that can be verified.** Panel (a) introduces the prediction performance of the model with the known cancer (breast and prostate cancer) related miRNAs. Panel (b) shows the prediction performance after the removal of the existing associations. The x-axis is the number of predictions (× 10) while the y-axis is the percentages of the verified predictions. . 60

4.1 **An example: From a DGR tripartite network to a co-functional miRNA pair.** The network in panel (a) contains known associations among the genes g1, g2, g3, g4, and g5, the diseases d1, d2, d3, and d4, and the miRNAs R1, R2, R3, and R4. In this example, miRNAs R2 and R3 are both associated with all the four diseases. However, the other three miRNAs are each associated with only one of these diseases. All these four diseases are associated with two common genes g4 and g5. Meanwhile, both of g4 and g5 are the targets of miRNAs R2 and R3. It is believed that R2-R3-g4-g5 may form a functional module that associated with the development of all the four diseases. 66

4.2 **The flowchart of our prediction and scoring method.** Our work includes the parts such as material collection, similarity computing, association prediction, network reconstruction, scoring and priorization of the co-function miRNA pairs and result output. 67

4.3 **The 50 top-ranked co-functional miRNA pairs from the reconstructed cancer-miRNA-gene network.** The labels along the edges illustrate the co-function information of the miRNAs. The first number of each label is the rank of the corresponding pair according to our prioritization method. The following gene symbols are the validated common targets during the co-functioning of the pair of miRNAs. The last number shows the potential diseases that related to this co-function pair. The pair miR-195-5p-miR-15b-5p and the pairs formed by miR-29a/b/c-3p are highlighted and used as the examples to explain their co-function. 73

4.4 **The miR-29a-miR-29b-miR-29c co-function module, their targets and the enrichment analysis of the KEGG pathways.** The triangles are the potential common target genes of the miR-29a/b/c co-functional module. Those small squares are the genes enriched pathways. Those disease names in the big squares are the co-functional module related diseases according to our prioritization method. 78

5.1 **The flowchart for the vectorization representation of a disease-lncRNA gene pair.** A disease-lncRNA gene pair can be represented by the integration of four sub-vectors including disease gene chromosome substructures' distribution information entropy vector (disease gene distribution vector), the disease gene enriched pathway groups' distribution information entropy vector (disease pathway distribution vector), the lncRNA gene sequence's k-mer frequency vector and the lncRNA gene expression profile. 86

5.2 **The disease chromosome enrichment analysis pie graph.** Subchr means chromosome substructure. We did the statistics of how many chromosomes a disease gene set enriches. More than a half (53%) of the 2802 diseases are just enriched to only one chromosome substructure, while just 3% of these diseases can be enriched to more than 4 chromosome substructures. . . 96

5.3 **The disease chromosome enrichment analysis results.** The y-axis are percentages of diseases that enriched to each of the chromosome substructures. The x-axis are the indexes of the chromosome substructures. The bar graph at the top right shows the statistics of the numbers of chromosome substructures that contained by diseases with given percentages scopes. 97

5.4 **The ROC curves of different methods for computing the disease similarities.** There are 7 ROC curves: the disease pathway distribution entropy vector method (θ=0, AUC=0.8555); the disease gene distribution entropy vector method (θ=1, AUC=0.9067); the integrated similarity method (θ=0.8, AUC=0.9094); the pathway status series vector method (AUC=0.7867); the disease gene status series vector method (AUC=0.5882); FunSim (AUC=0.8858) and Symptom representation method (AUC=0.7455). 100

5.5 **The boxplot graph of the AUC values for the 5-fold cross validation experiments.** The x-axis is the value of V, and the y-axis is the corresponding AUC values. The changes of the AUC values with different V are tiny. For a given V, the prediction results are stable. 104

5.6 **The leave-one-out cross validation results based on three datasets with different methods.** Four methods were compared, our method with type 7 (W=7) feature and type 1 (W=1) feature, LRLSLDA method and the LRLSLDA_ILNCSIM method. Our type 7 method works best for all three datasets. 105

5.7 **The final prediction test on the lncRNADisease dataset.** The x-axis is the unknown disease-lncRNA pairs' predicted ranks. The y-axis are the predicted scores which means the possibilities of the samples to be positive. The predicted results were validated via the lnc2cancer and MNDR datasets. The validated samples were marked on the score curve. The ROC curve that compares the scores of the validated samples and the remain unknown samples is drawn at the top right of this figure. The AUC value achieves 0.9005. 106

6.1 **The flowchart to construct TSAM for predicting sgRNA cleavage efficiencies.** This flowchart contains four main steps: at first 6 types of initial features are created; in the second step, primary features are selected from the initial feature set to optimize an XGBoost regressor and output the first-step scores (fss) and the importance scores of the features; then, the important features are combined with the pHMM features to train an RBF kernel SVM and compute the second-step score (sss); lastly, the first-step score and the second-step score of a sgRNA is averaged as the final predicted score (($fss + sss$)/2). 120

6.2 **Top 12 important features and analysis on the nucleotide and cleavage preferences.** Y-axis shows the feature values. The feature names are placed under the x-axis and their symbols are placed at the top right panel of the subplots. These features are ranked by their importance. Type “high” means that the sgRNAs are ranked at top-20% while the “low” represents that the sgRNAs are ranked at bottom-20%. The p-value shown in each sub-figure is computed via the two-sample Kolmogorov-Smirnov test. 122

7.1 **An example of on-target site and off-target sites. The on-target site is the expected binding site for an sgRNA.** The off-target sites are unintended binding sites and the off-target editing effect should be avoided in practical use. The spacer in the sgRNA is the RNA version of the protospacer sequence that is located in the genome DNA. Sometimes the spacer and protospacer are interchangeably used. The protospacer sequence determines where for the sgRNA to bind, and the existence of a protospacer adjacent motif (PAM) determines whether it cuts at the target site. . . 133

7.2 **An example of a sequence pair $\langle onTSeq, offTSeq \rangle$.** The mismatches are those pairs of nucleotides at the given position but with different nucleotide type such as at the positions of 4, 11 and 12 140

7.3 **Comparison of the mismatch distributions in the positive and negative sample sets.** The lines depict the remarkable distribution differences between the two groups. . . 147

7.4 **Receiver Operating Characteristic curves (left) and Precision-Recall curves (right) for the cross-dataset validation of our proposed method and the four state-of-the-art methods.** 148

7.5 **Overlap rates of different computation methods relative to the high-throughput sequencing base methods.** The proposed method detected off-targets overlaps better than other computational methods relative to all the sequencing methods' results. Sequencing methods CIRCLE, Digenome, GUIDE, HTGTS and mDigenome refers to the CIRCLE-seq, Digenome-seq, GUIDE-seq, HTGTS, multiplex Digenome-seq. The 'Integrated' means the union result of the four sequencing methods. 152

List of Tables

2.1 **The existing tools for CRISPR/Cas9 on-target cutting efficiency prediction.** 27

2.2 **The existing off-target site detection methods.** 32

3.1 **The prediction performances based on different approaches to select negative samples** 55

3.2 **Performance comparison between our method and the three state-of-the-art prediction methods.** Symbols "+/-" represent "positive samples/negative samples". cv means cross-validation. 56

4.1 **The co-functional miRNA pairs and their potential co-functional targets for both cancers and non-cancer diseases** 76

5.1 **Feature types and their corresponding performance.** . 102

5.2 **Case studies for predicting breast cancer and prostate cancer related lncRNAs.** 107

6.1 **11 datasets for construction and evaluation of our classification and regression models** 114

6.2 **Regression performance of different methods on four benchmark datasets.** 125

6.3 Performance comparison between our method and the state-of-the-art methods for the binary classification of sgRNAs. 127
6.4 Spearman correlation of TSAM, RS2 and CRISPRscan tested on datasets from U6 or T7 expression systems. 128

7.1 The datasets for constructing the positive sample sets. 138
7.2 AUROC and AUPRC scores of the proposed method and the state-of-the-art methods in various tests. . . . 149
7.3 The ranks of the sgRNAs by considering both of their cutting efficiencies and off-target potentials. 153

A.1 The example 2∗2 contingency table 164

List of Publications

Below is the list of journal and conference papers associated with my PhD research:

Journal Papers Published

- **Peng, H.**, Zheng, Y., Blumenstein, M., Tao, D., & Li, J. (2018). CRISPR/Cas9 cleavage efficiency regression through boosting algorithms and Markov sequence profiling. **Bioinformatics**, 34(18), pp.3069-3077.

- **Peng, H.**, Zheng, Y., Zhao, Z., Liu, T., & Li, J. (2018). Recognition of CRISPR/Cas9 off-target sites through ensemble learning of uneven mismatch distributions. **Bioinformatics**, 34(17), pp.i757-i765. (Oral Presentation at the 17th Europe Conference on Computational Biology (ECCB 2018))

- Zheng, Y., **Peng, H.**, Ghosh, S., Lan, C., & Li, J. (2018). Inverse Similarity and Reliable Negative Samples for Drug Side-effect Prediction. **BMC Bioinformatics**, 19(13), p.554.

- Zheng, Y., **Peng, H.**, Zhang, X., Zhao, Z., & Li, J. (2018). Predicting adverse drug reactions of combined medication from heterogeneous pharmacologic databases. **BMC Bioinformatics**, 19(19), p.517.

- Lan, C., **Peng, H.**, McGowan, G., Hutvagner, G., & Li, J. (2018). An isomiR expression panel based novel breast cancer classification approach using improved mutual information. **BMC Medical Genomics**, 11(6), p.118.

- Ho, N., **Peng, H.**, Mayoh, C., Liu, P. Y., Atmadibrata, B., Marshall, G. M., ... & Liu, T. (2018). Delineation of the frequency and boundary of chromosomal copy number variations in paediatric neuroblastoma. **Cell Cycle**, 17(6), pp.749-758. **(co-first author)**

- Zhao, Z., **Peng, H.**, Lan, C., Zheng, Y., Fang, L., & Li, J. (2018). Imbalance learning for the prediction of N6-Methylation sites in mRNAs. **BMC Genomics**, 19(1), p.574.

- **Peng, H.**, Lan, C., Liu, Y., Liu, T., Blumenstein, M., & Li, J. (2017). Chromosome preference of disease genes and vectorization for the prediction of non-coding disease genes. **Oncotarget**, 8(45), p.78901.

- **Peng, H.**, Lan, C., Zheng, Y., Hutvagner, G., Tao, D., & Li, J. (2017). Cross disease analysis of co-functional microRNA pairs on a reconstructed network of disease-gene-microRNA tripartite. **BMC Bioinformatics**, 18(1), p.193.

- Liu, Y., **Peng, H.**, Wong, L., & Li, J. (2017). High-speed and high-ratio referential genome compression. **Bioinformatics**, 33(21), pp.3364-3372.

Conference Papers

- Zheng, Y., **Peng, H.**, Zhang, X., Gao, X., & Li, J. (2018). Predicting Drug Targets from Heterogeneous Spaces using Anchor Graph Hashing and Ensemble Learning. **International Joint Conference on Neural Networks**.

Papers to be Submitted/Under Review/Accepted

- Liu, P., Tee, A., Milazzo, G., Hannan, K., Maag, J., Mondal, S., Atmadibrata, B., Bartonicek, N., **Peng, H.**, Ho, N., Mayoh, C., Sun, Y., Welham, Z., Hulme, A., Henderson, M., Wong, M., Lan, Q., Cheung, B., Wang, J., Simon, T., Fischer, M., Zhang, X., Marshall, G., Norris, M., Haber, M., Vandesompele, J., Li, J., Mattick, J., Mestdagh, P., Hannan, R., Dinger, M., Perini, G., & Liu, T. (2018). The novel long noncoding RNA lncNB1 promotes tumorigenesis by interacting with ribosomal protein RPL35. **Nature Communications**.

- Lan, C., **Peng, H.**, Hutvagner, G., & Li, J. (2018). Construction of Competing Endogenous RNA Networks from Paired RNA-seq Data Sets by Pointwise Mutual Information Theories. **Bioinformatics**. (Major revision)

Abstract

This thesis presents computational methods used for the recognition of disease genes and for the optimal design of disease gene CRISPR/Cas9 editing systems. The key innovation in these computational methods is the feature space and characteristics captured from the biology domain knowledge through machine learning algorithms.

The disease-gene association prediction problems are studied in Chapters 3-5. Disease gene recognition is a hot topic in various fields, especially in biology, medicine and pharmacology. Non-coding genes, a type of genes without protein products, have been proved to play important roles in disease development. Particularly, the two kinds of non-coding gene products such as microRNA (miRNA) and long non-coding RNA (lncRNA) have caught much attention as they are abundantly expressed in various tissues and frequently interact with other biomolecules, e.g. DNA, RNA and protein. The disease-ncRNA relationships remain largely unknown. Computational methods can immensely help replenish this kind of knowledge. To overcome existing computational methods' limitations such as significantly relying on network structures and similarity measurements, or lacking reliable negative samples, this thesis presents two novel methods.

One is the precomputed kernel matrix support vector machine (SVM) method to predict disease related miRNAs in Chapter 3. The precomputed kernel matrix was built by integrating several kinds of similarities computed with effective characteristics for miRNAs and diseases. The reliable negative samples were collected through analyzing the published array and sequencing data. This binary classification method accurately predicts disease-miRNA associations, which outperforms those state-of-the-art methods. In Chapter 4, the predicted novel disease-miRNA associations were combined with known relationships of diseases, miRNAs and genes to reconstruct a disease-gene-miRNA (DGR) tripartite network. Reliable multi-disease associated co-functional miRNA pairs were extracted from this DGR for cross-disease analysis by defining the co-function score. This not only proves the proposed method's effectiveness but also contributes to the study of multi-purpose miRNA therapeutics.

Another is the bagging SVM-based positive-unlabeled learning method for disease-lncRNA prioritizing that is described in Chapter 5. It creatively characterized a disease with its related genes' chromosome distribution and pathway enrichment properties. The disease-lncRNA pairs were represented as novel feature vectors to train the bagging SVM for predicting disease-lncRNA associations. This novel representation contributes to the superior performance of the proposed method in disease-lncRNA prediction even when a given disease has no currently recognized lncRNA genes.

After confirming the relationships between genes and diseases, one of the most difficult tasks is to investigate the molecular mechanism and treatment of the diseases considering their related genes. The CRISPR/Cas9 system is a promising gene editing tool for operating the genes to achieve the goals of disease-gene function clarification and genetic disease curing. Designing an optimal CRISPR/Cas9 system can not only improve its editing efficiency but also reduce its side effect, i.e. off-target editing. Furthermore, the off-target site detection problem involves genome-wide sequence observing which makes it a more challenging job. The CRISPR/Cas9 system on-target cutting efficiency prediction and off-target site detection questions are discussed in Chapters 6 and 7 respectively.

To accurately measure the CRISPR/Cas9 system's cutting efficiency, the profiled Markov properties and some cutting position related features were merged into the feature space for representing the single-guide RNAs (sgRNAs). These features were learned by a two-step averaging method where an XGBoost's predictions and an SVM's predictions were averaged as the final results. Later performance evaluations and comparisons demonstrate that this method can predict a sgRNA's cutting efficiency with consistently good performance no matter it is expressed from a U6 promoter in cells or from a T7 promoter in vitro.

In the off-target site detection, a sample was defined as an on-target-off-target site sequence pair to turn this problem into a classification issue. Each sample was numerically depicted with the nucleotide composition change features and the mismatch distribution properties. An ensemble classifier was constructed to distinguish real off-target sites and no-editing sites of a given sgRNA. Its excellent performance was confirmed with different test scenarios and case studies.