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

# List of Figures

# List of Tables

# 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.

# Chapter 1

# Introduction

This chapter introduces the background knowledge of the research topics in Section 1.1. Then, the research questions and their formulations are presented in Section 1.2. The contributions and structure of this thesis are described in Sections 1.3 and 1.4 respectively.

## 1.1 Background

This thesis introduces two topics of my research such as disease gene (mainly non-coding gene) recognition and optimal design of CRISPR/Cas9 system for disease gene editing. In this section, protein-coding genes and non-coding genes, disease-non-coding RNA associations, the CRISPR/Cas9 system and machine learning related background knowledge are described.

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

Genes are always referring to the functional regions in the genome DNA (Portin & Wilkins 2017). The gene expression includes a transcription process where the gene information is transcribed into a corresponding RNA, e.g. a messenger RNA (mRNA) for a protein-coding gene or a non-coding RNA (ncRNA) for a non-coding gene (Alberts, Johnson, Lewis, Walter, Raff &

Roberts 2002). The mRNA is then translated to produce the protein while those ncRNAs are not. It was estimated that about 75% of the human genome can be transcribed to RNAs. Among these RNAs, just 3% of them have the protein coding ability (Ling, Fabbri & Calin 2013). Currently, the functions of those protein-coding genes have been widely investigated as their protein products are stable enough to be observed and experimentally validated. In comparison, clarifying the non-coding genes' functions is more difficult because of the ncRNAs' instability and diversity (Eddy 2001). The ncRNAs transcribed from the non-coding genes include ribosomal RNAs (rRNAs), small nucleolar RNAs (snoRNAs), microRNAs (miRNA), long non-coding RNAs (lncRNAs) and so on (Palazzo & Lee 2015). Among them, the miRNAs and lncRNAs have attracted increasing attention as they were reported to contain significant genetic information and functions (Mattick & Makunin 2006, Derrien, Johnson, Bussotti, Tanzer, Djebali, Tilgner, Guernec, Martin, Merkel, Knowles et al. 2012).

The miRNA is a type of evolutionarily conserved small RNA with ∼22 nucleotides, which is produced by the RNase-III-type enzyme Dicer from an endogenous transcript that contains a local hairpin structure (Ambros, Bartel, Bartel, Burge, Carrington, Chen, Dreyfuss, Eddy, Griffiths-Jones, Marshall et al. 2003, Kim 2005, Bartel 2018). The miRNAs were estimated to account for 1-5% of the human genome and more than 30% of protein-coding genes are regulated by them (Berezikov, Guryev, van de Belt, Wienholds, Plasterk & Cuppen 2005, Rajewsky 2006, MacFarlane & R Murphy 2010). It was also reported that there may be about 40% of miRNA genes located in the introns or even exon regions of other genes (Rodriguez, Griffiths-Jones, Ashurst & Bradley 2004, Baskerville & Bartel 2005, Kim & Kim 2007). The miRNAs regulate their target mRNAs' expressions at the post-transcriptional level by RNA degradation or translation repression (Eulalio, Huntzinger, Nishihara, Rehwinkel, Fauser & Izaurralde 2009, Wahid, Shehzad, Khan & Kim 2010, Gebert & MacRae 2018). These processes are completed via pairing up with the complementary sequences within their target

mRNAs (Bartel 2009).

The lncRNA is another type of ncRNA with more than 200 nucleotides. LncRNAs can be grouped into five categories according to their genomic locations such as stand-alone lncRNAs, i.e. large intergenic (or intervening) ncRNAs (lincRNA), antisense transcripts, pseudogenes, long intronic ncRNAs and others (Kung, Colognori & Lee 2013). Different from those small ncRNAs, most (81%) lncRNAs are poorly conserved (Wang & Chang 2011, Fang & Fullwood 2016). However, the lncRNAs also play key roles in gene expression regulation via various actions such as gene activation or repression (Carpenter, Aiello, Atianand, Ricci, Gandhi, Hall, Byron, Monks, Henry-Bezy, Lawrence et al. 2013), acting as miRNA sponges (Hansen, Jensen, Clausen, Bramsen, Finsen, Damgaard & Kjems 2013), regulating mRNA degradation (Liu, Li, Zhang, Guo & Zhan 2012, Dykes & Emanueli 2017) and others (Wang & Chang 2011, Kung et al. 2013, Marchese, Raimondi & Huarte 2017).

### 1.1.2 Non-coding RNA and human diseases

The dysregulation of ncRNAs may result in the aberrant expression of their target genes or disturbing their related cellular processes, which finally cause the development of various diseases. For example, the miRNA miR-21 was proved to be a regulator of the ERKMAP kinase signaling pathway, which relates to the myocardial diseases (Thum, Gross, Fiedler, Fischer, Kissler, Bussen, Galuppo, Just, Rottbauer, Frantz et al. 2008). Some researchers drew the conclusion that loss of miRNA-29a/b-1 cluster may result in the over-expression of BACE1/$\beta$-secretase gene and finally associate with the causing of Alzheimer's disease (Hébert, Horré, Nicolaï, Papadopoulou, Mandemakers, Silahtaroglu, Kauppinen, Delacourte & De Strooper 2008). The lncRNA-NUTF2P3-001 was experimentally verified to be significantly over-expressed in both the pancreatic cancer and the chronic pancreatitis tissues (Li, Deng, Zhu, Jin, Cui, Chen, Xiang, Li, He, Zhao et al. 2016). Loc285194, another lncRNA, was also proved to be a tumor suppressor that

regulates p53 (Liu, Huang, Zhou, Zhang, Zhang, Lu, Wu & Mo 2013).

Disease development describes the process of appearing disorders that affect the normal condition of an organism. Here, we just study the disease development associated with the dysfunction of genes that dys-regulated by other functional molecules during different cell processes but not the physical injury or virus infection.

During the past decade, some databases have collected abundant of verified associations between the miRNAs or lncRNAs and diseases. The HMDD (Li, Qiu, Tu, Geng, Yang, Jiang & Cui 2013) and miR2Disease (Jiang, Wang, Hao, Juan, Teng, Zhang, Li, Wang & Liu 2009) are two of the most popular databases that store the experimentally verified disease-miRNA associations. The validated disease-lncRNA associations can be found from the databases LncRNADisease (Chen, Wang, Wang, Qiu, Liu, Chen, Zhang, Yan & Cui 2013) and Lnc2Cancer (Ning, Zhang, Wang, Zhi, Wang, Liu, Gao, Guo, Yue, Wang et al. 2016). However, comparing with the known species of ncRNAs and diseases, the known associations of them are rare. Traditional biological experiments for finding the associations are time-consuming and expensive. Thus, over the past few years, many computational prediction methods have been designed to predict the potential disease-associated ncRNAs at a large scale. They require only known associations and other related information as inputs but achieve good prediction performances.

### 1.1.3 The CRISPR/Cas9 system

Gene editing (or genome editing) is the technique to change an organism's genome DNA. Previously reported gene editing tools include the zinc finger nucleases (ZFNs) (Urnov, Miller, Lee, Beausejour, Rock, Augustus, Jamieson, Porteus, Gregory & Holmes 2005, Urnov, Rebar, Holmes, Zhang & Gregory 2010) and the transcription activator-like effector nucleases (TALENs) (Reyon, Tsai, Khayter, Foden, Sander & Joung 2012, Joung & Sander 2013). Recently, a third generation gene editing tool, the Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR-associated protein

9 (CRISPR/Cas9) was generated (Ran, Hsu, Wright, Agarwala, Scott & Zhang 2013, Shalem, Sanjana, Hartenian, Shi, Scott, Mikkelsen, Heckl, Ebert, Root, Doench et al. 2014, Wang, Wei, Sabatini & Lander 2014, Doudna & Charpentier 2014). The CRISPR/Cas9 system is thought to be faster, cheaper, more accurate and efficient than ZFNs and TALENs (Wang, Yang, Shivalila, Dawlaty, Cheng, Zhang & Jaenisch 2013, Xiong, Ding & Li 2015, Kanchiswamy, Sargent, Velasco, Maffei & Malnoy 2015). It is composed of a CRISPR associated protein 9 (Cas9) and a single-guide RNA (sgRNA) as a complex (see Figure. 1.1). The sgRNA always contains at least a CRISPR-RNA (crRNA) and a trans-activating crRNA (tracrRNA) which are linked by a linker loop. There is a 20 nucleotides (20nt) spacer sequence at the 5'-end of the crRNA which can pair-up with its target DNA region (Nishimasu, Ran, Hsu, Konermann, Shehata, Dohmae, Ishitani, Zhang & Nureki 2014, Jiang & Doudna 2017). The sgRNA can locate the CRISPR/Cas9 system to its target region by pairing up with the DNA sequence opposite to the protospacer. The Cas9 protein is an RNA-guided DNA endonuclease enzyme which has two function domains such as the HNH and RuvC (Jinek, Jiang, Taylor, Sternberg, Kaya, Ma, Anders, Hauer, Zhou, Lin et al. 2014, Anders, Niewoehner, Duerst & Jinek 2014, Nishimasu et al. 2014). This protein can recognize a 3nt protospacer adjacent motif (PAM), e.g. 'NGG', and cut the PAM upstreamed protospacer sequence and its complementary sequence (Anders et al. 2014).

The genome cleavage process begins with the locating of a CRISPR/Cas9 system to the region where one of its DNA sequences can pair-up with the spacer. Then, the Cas9 protein binds to this region and tries to recognize the downstream PAM. If a PAM exists, the double-strand break (DSB) is generated at the position 3nt upstream to this PAM by the Cas9 (Hsu, Lander & Zhang 2014, Doudna & Charpentier 2014). After cleavage being generated, the cell senses this problem and activates the repairing mechanisms such as homology-directed repair (HDR) and non-homology end joining (NHEJ) to repair this problem (Ran, Hsu, Lin, Gootenberg, Konermann, Trevino, Scott,

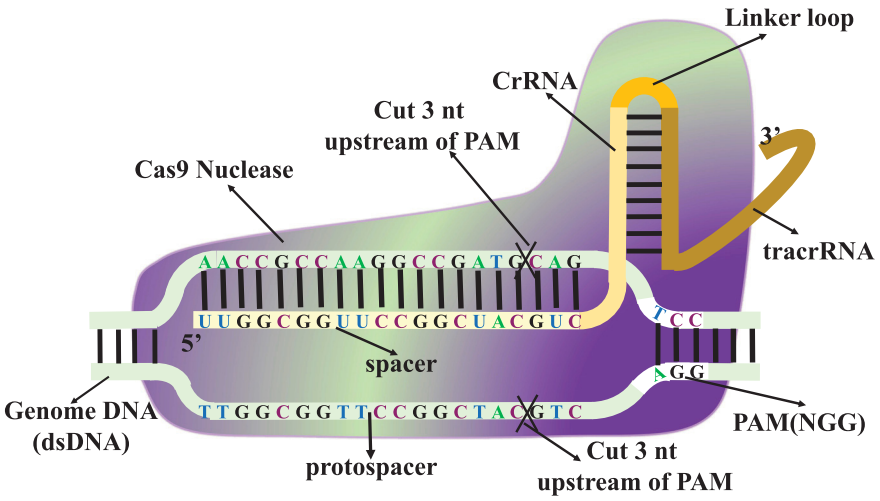


Figure 1.1: **An example of a CRISPR/Cas9 system cutting a genome DNA sequence.**

Inoue, Matoba, Zhang et al. 2013, Kan, Ruis, Takasugi & Hendrickson 2017). During the repairing process, one can insert a fragment into the genome via HDR, or knock-out a fragment with NHEJ to complete the genome editing task.

### 1.1.4 CRISPR/Cas9 system design for disease gene editing

As is described above, the target regions of a CRISPR/Cas9 system have two properties: existing a protospacer where its complementary sequence can pair-up with the spacer and existing a PAM which is downstream to the protospacer. After injecting a CRISPR/Cas9 system into the cell, if it generates the DSB at our expected target region, then we call this action an on-target cutting. However, if this system cuts the genome at unwanted regions, off-target cuttings will happen. The CRISPR/Cas9 cutting at an unintended region is mainly due to the fact that the 20nt protospacer + 3nt

PAM sequence may not uniquely exist in the genome. Especially, during the base-pairing of a spacer and the DNA sequence, mismatches or even bulges are permitted (Fu, St Onge, Fire & Smith 2016, Lee, Cradick & Bao 2016). There may be thousands of such kind of 23nt sequences in the whole genome when as many as 6 or more mismatches are permitted. Thus, off-target cuttings may possibly exist for a designed CRISPR/Cas9 system, which can result in serious toxic effects.

The CRISPR/Cas9 system is now one of the most widely applied gene editing tools in various fields such as gene function investigation (Swiech, Heidenreich, Banerjee, Habib, Li, Trombetta, Sur & Zhang 2015), disease model construction (Platt, Chen, Zhou, Yim, Swiech, Kempton, Dahlman, Parnas, Eisenhaure, Jovanovic et al. 2014) and genetic disease treatment (Wu, Liang, Wang, Bai, Tang, Bao, Yan, Li & Li 2013). Disease-gene editing helps reveal the role of this gene played in the disease development and benefits the accurate modeling and treating of the disease. A corresponding CRISPR/Cas9 system is required to be designed for editing a given disease gene. An optimal design of the CRISPR/Cas9 system is that its on-target cutting efficiency is high and it has no off-target effect. However, fully avoiding the off-target effect is extremely hard. Designing the system with higher on-target cutting efficiency but with lower off-target effect is an alternative goal in practical usage. The design of a CRISPR/Cas9 system in this thesis means to select a 20nt spacer sequence for the sgRNA. In following contents, we may call the CRISPR/Cas9 system design task a sgRNA selection process or a spacer choosing step instead. In addition, for convenience, the statement of "cutting efficiency of a spacer or a sgRNA" equals to "cutting efficiency of the CRISPR/Cas9 system".

### 1.1.5 Machine learning in bioinformatics

Machine learning algorithms are powerful tools for addressing many bioinformatics problems such as molecular structure prediction (Ward, McGuffin, Buxton & Jones 2003, Bindewald & Shapiro 2006), molecular function

prediction (Cai, Han, Ji, Chen & Chen 2003, Barutcuoglu, Schapire & Troyanskaya 2006, Schietgat, Vens, Struyf, Blockeel, Kocev & Džeroski 2010) and association prediction (Bock & Gough 2001, Liu, Wu, Wang, Zhang & Chen 2010, Zhang, Zhou, Hu, Gong, Chen, Cheng & Zeng 2015). The application of a machine learning algorithm usually contains a model training process. According to whether labels are available during the model training, machine learning algorithms can be classified into supervised learning (Kotsiantis, Zaharakis & Pintelas 2007), semi-supervised learning (Chapelle, Scholkopf & Zien 2009) and unsupervised learning (Hastie, Tibshirani & Friedman 2009), which correspond to labels provided, part of the labels provided and no label provided respectively. Machine learning algorithms are always used to solve classification, regression or clustering problems. In a classification problem, a sample can be classified to belong to one of the known categories. However, in a regression question, a continuous value is assigned to this sample. For a clustering issue, the input samples are clustered into several groups, where no group information is known before the learning (usually learned by unsupervised algorithms).

Many bioinformatics problems can be transformed into classification, regression or clustering issues. For example, the disease-gene prediction can be regarded as a classification problem, where a gene can be labeled as a disease gene or a non-disease gene (Le, Hoai & Kwon 2015). The CRISPR/Cas9 system cutting efficiency measurement can be solved by regression methods (Doench, Fusi, Sullender, Hegde, Vaimberg, Donovan, Smith, Tothova, Wilen, Orchard et al. 2016). There are two fundamental steps when applying machine learning to solve bioinformatics questions. One is the collection of reliable training samples and another is the feature space construction for sample representation. As the reliable samples are always collected from biological experiments, obtaining enough samples for building effective models are always difficult. Designing excellent methods to make good use of the limited samples for optimizing the final model is also a significant task. On the other hand, the construction of feature spaces is

important as good features can significantly improve the final performance of the model (Guyon & Elisseeff 2003, Saeys, Inza & Larrañaga 2007). With the samples and feature spaces being prepared well, the following step is to select an appropriate machine learning algorithm. The widely applied algorithms include decision tree (Safavian & Landgrebe 1991), support vector machine (SVM) (Hearst, Dumais, Osuna, Platt & Scholkopf 1998), random forest (Liaw, Wiener et al. 2002) and deep learning algorithms (LeCun, Bengio & Hinton 2015) and so on. Later, the model can be optimized via parameter tuning with cross-validations (Kohavi 1995). The final model can be provided to users for completing new prediction tasks.

## 1.2 Research Questions and Formulations

This thesis mainly discusses two research topics such as disease-ncRNA gene prediction and disease gene editing tool design. There are several specific questions need to be solved for these two topics. The detail research questions and their formulations are listed in below Q1 to Q3.

- **Q1: Disease-ncRNA gene association prediction**

  This thesis only focuses on two kinds of ncRNA genes such as miRNA and lncRNA. For a given disease $d$ and a ncRNA gene $r$, this research question can be simply expressed as the following formula:

$$f(d,r) = \begin{cases} 1 & if\ r\ relates\ to\ d, \\ 0 & else. \end{cases} \tag{1.1}$$

  Here, $f$ is a well-trained binary classification model to classify a disease-ncRNA pair $(d, r)$ as positive (1, if the ncRNA relates to the disease) or negative (0, otherwise). There are four kinds of difficulties in solving this question: (1) Lacking negative samples. The existing databases only collected the validated disease-ncRNA associations but neglected negative ones (i.e., no association between a disease and a

ncRNA); (2) NcRNA and disease representation. The RNA sequences have different lengths and their properties vary widely. Especially, diseases are phenotypes, their mathematical characterization seems extremely difficult; (3) Outperforming the existing methods. A number of computational methods have been proposed for solving this question. Overcoming their limitations and improving the prediction performances are challenging; (4) The validation of outputs. We cannot conduct wet-lab experiments to validate our newly predicted associations.

- **Q2: CRISPR/Cas9 system on-target cutting efficiency prediction**

  For a given disease gene $g$ (coding or non-coding gene) to be cut by the CRISPR/Cas9 system, the cutting efficiency $ce_i$ of its candidate sgRNA $s_i$ is predicted via the below function:

$$f\left(s_i|g\right) = ce_i \tag{1.2}$$

  The function $f$ is an optimized regression model. The output efficiency $ce_i$ is a continuous value between 0 and 1 where 1 represents the highest cutting efficiency. The challenges for addressing this question are: (1) New feature space is required to be constructed for representing sgRNAs. Currently adopted features may lose valuable information. More domain knowledge based characteristics should be extracted; (2) The prediction performance needs to be improved. There exists large space for improving the prediction performances of the state-of-the-art methods. Novel prediction strategy needs to be applied; (3) Investigating the effects of expression systems and species on the cutting efficiency prediction. The existing datasets are collected from various expression systems (e.g. U6 or T7) and species (e.g. mammals and zebrafish). The data source's effect on the prediction performance

has not been clearly investigated. (4) Web server construction. A web server is necessary for helping users access to the proposed tool.

- **Q3: CRISPR/Cas9 system off-target site detection**

  Each of the given sgRNA $s_i$ for cutting a disease gene $g$ has $N$ candidate editing sites $O = \{o_1, o_2, \cdots, o_j, \cdots, o_N\}$, we need to label $o_j$ with the following model:

$$f(o_j|s_i, g) = \begin{cases} 1 & if\ o_j\ is\ edited, \\ 0 & else. \end{cases} \tag{1.3}$$

  $f$ is a well-learned classifier to determine whether a given candidate site $o_j$ is an off-target site (being edited) or not. The troubles to be handled include: (1) Transforming the off-target site detection problem into a classification issue. Simply defining a candidate editing site as the sample to be classified is not reasonable; (2) Data collection. The already known sgRNAs' off-target sites are rare. The experimental platforms for finding them are different. It is hard to collect and clear those heterozygous data to generate reliable datasets; (3) Imbalanced learning. Each designed sgRNA contains thousands of potential editing sites. Only a small part of them may be real off-target sites. The training dataset must be extremely imbalanced; (4) Performance evaluation and comparison. The size of those reliable data is small. Splitting the data reasonably for optimizing and validating the prediction model is difficult. In addition, comparing the proposed machine learning method with those existing methods including the wet-lab technologies and the computational methods is uneasy.

## 1.3 Research Contributions

Corresponding to the above research questions, we have proposed novel methods to solve them (described in Chapters 3 to 7). Our contributions

are concluded as the following C1 to C5.

- **C1: Precomputed kernel matrix SVM for disease-related miRNA prediction**
  Chapter 3 presents a precomputed kernel matrix SVM method for predicting disease-related miRNAs. The contributions of this part of work are four-fold: (1) New strategy was adopted to select reliable negative samples. We regarded those non-significantly differentially expressed miRNAs as non-disease related miRNAs. We then turned the disease-miRNA association prediction problem into a binary classification issue; (2) A precomputed kernel matrix was designed as the input of SVM instead of traditional features. We used new characteristics to measure the disease similarities and miRNA similarities for constructing the precomputed kernel matrix. There is no need to do feature selection and SVM parameter tuning during the prediction model construction; (3) The proposed method achieves better performance than the existing state-of-the-art methods. This was confirmed by various cross-validation tests; (4) By applying this model, some novel disease-miRNA associations were predicted and validated. For example, the predicted breast cancer-hsa-miR-15b and prostate cancer-hsa-miR-29c have been validated by newer literature.

- **C2: Cross disease analysis of co-functional miRNA pairs on a reconstructed network of disease-gene-miRNA tripartite**
  In Chapter 4, we describe an extended study about the application of disease-miRNA associations, where the co-functional roles of miRNA pairs in multi-disease development are explored. Our contributions are: (1) By combining our predicted novel disease-related miRNAs with existing disease-miRNA associations, miRNA-target gene relationships and disease-target gene associations, we reconstructed a disease-gene-miRNA tripartite network (DGR); (2) We defined the multi-disease related co-functional miRNA pair co-function score. By this definition, the reliable multi-disease related co-functional miRNA pairs can be

prioritized for further investigation; (3) Reliable multi-disease related co-functional miRNA pairs were found. For example, we found that the pair miR-15b-miR-195 may contribute to the development of as many as 38 different cancers through dys-regulating their target gene BCL2. These multi-disease related miRNA pairs can help for multi-propose drug design; (4) From the cross-disease analysis of co-functional miRNAs, we found that the co-function phenomenon is not unusual. We also confirmed that the regulations of miRNAs for the development of cancers are more complex and have more unique properties than those of non-cancer diseases.

- **C3: Chromosome preference of disease genes and vectorization for the prediction of non-coding disease genes**
  Chapter 5 describes the proposed positive-unlabeled learning method based on the bagging SVM to prioritize disease related lncRNAs. Our contributions include: (1) We found that there exist distribution preferences of disease-genes on the chromosomes. One is that disease-genes of a given disease are very likely located at a neighborhood region. Another is that disease genes are unevenly located on the chromosomes, e.g. p-arm of chromosome 6 is the most preferred substructure of disease genes - about 16.2% of the disease-related gene sets can be enriched here; (2) We characterized diseases with two types of novel vectors such as the disease-gene chromosome substructures' distribution information entropy vector and the disease-gene enriched pathway groups' distribution information entropy vector. With these two vectors, we can accurately compute the similarity between two diseases. We also characterized the lncRNAs with feature vectors on the basis of their sequence and expression profile information. (3) A bagging SVM positive-unlabeled learning method was adopted to handle the difficulty of lacking negative samples; (4) With the comparisons and case studies, we proved the excellent performance of our proposed method on the disease-lncRNA association prediction.

- **C4: CRISPR/Cas9 cleavage efficiency regression through boosting algorithms and Markov sequence profiling**
  In Chapter 6, a CRISPR/Cas9 system's cutting efficiency prediction tool is presented. This part work contributes to several aspects: (1) We extracted new domain knowledge characteristics to represent sgRNAs. These new features include the profiled hidden Markov properties, some sequence composition features and the cutting position features; (2) We proposed a two-step average method for the prediction of CRISPR/Cas9 on-target cutting efficiency. We carried out the first step prediction with an XGBoost framework. Then, a second round prediction was implemented via learning the XGBoost selected important features. The two rounds of predictions were averaged as the final predicted scores; (3) The applied prediction strategy significantly improved the performance comparing to the state-of-the-art methods; (4) We found some important features that affect the cutting efficiency largely such as the cutting position relative to not only the protein but also the transcript; (5) We confirmed that the sgRNA expression system affects the efficiency prediction. A number of datasets were used to conduct the tests for this confirmation; (6) A web server tool was provided which can predict the cutting efficiency of a sgRNA no matter it is expressed from a U6 or a T7 promoter; (7) Two case studies validated the good performance of the proposed method for designing sgRNAs to cure genetic diseases in practical use.

- **C5: Recognition of CRISPR/Cas9 off-target sites through ensemble learning of uneven mismatch distributions**
  The Chapter 7 introduces a novel method for detecting off-target sites of the CRISPR/Cas9 system. Following outcomes have been obtained: (1) We turned the off-target site detection problem into a binary classification issue by defining a sample as a sequence pair that is composed of a given on-target site sequence and a corresponding candidate editing site sequence. The sequence pair

can be labeled as positive if its candidate editing site is a real off-target site, otherwise negative; (2) Samples were characterized with the nucleotide composition change and mismatch distribution properties. We observed significant differences of these two kinds of properties between positive samples and those negative ones; (3) We proposed an ensemble SVM classifier to detect off-target sites, which outperforms the existing computational methods. Our method also predicted more off-target sites that have been detected by the wet-lab technologies; (4) We applied two case studies to display the reliability of our method for sgRNA design in practical usage.

## 1.4 Thesis Structure

The structure of this thesis is shown in Figure. 1.2 and is simply explained below:

**Chapter 1** mainly introduces the background knowledge related to the work in this thesis. The research questions and our contributions are also described. **Chapter 2** reviews the related work about the studied problems in this thesis. For each of the three research questions such as disease-ncRNA prediction (including miRNA and lncRNA), CRISPR/Cas9 on-target cutting efficiency prediction and CRISPR/Cas9 off-target site detection, a section is used to present the current research progress about it. The limitations of existing methods and the inspirations from them are discussed at last. **Chapter 3** to **Chapter 7** describe the proposed methods for solving the research questions such as disease-miRNA prediction, multi-disease related co-functional miRNA pair extraction, disease-lncRNA prioritization, CRISPR/Cas9 on-target cutting efficiency prediction and CRISPR/Cas9 off-target site detection. These methods' construction, evaluation, comparison and related case studies are presented. Simple conclusions are given at the end of each chapter. **Chapter 8** summaries the work in this thesis and discusses my future research plans.

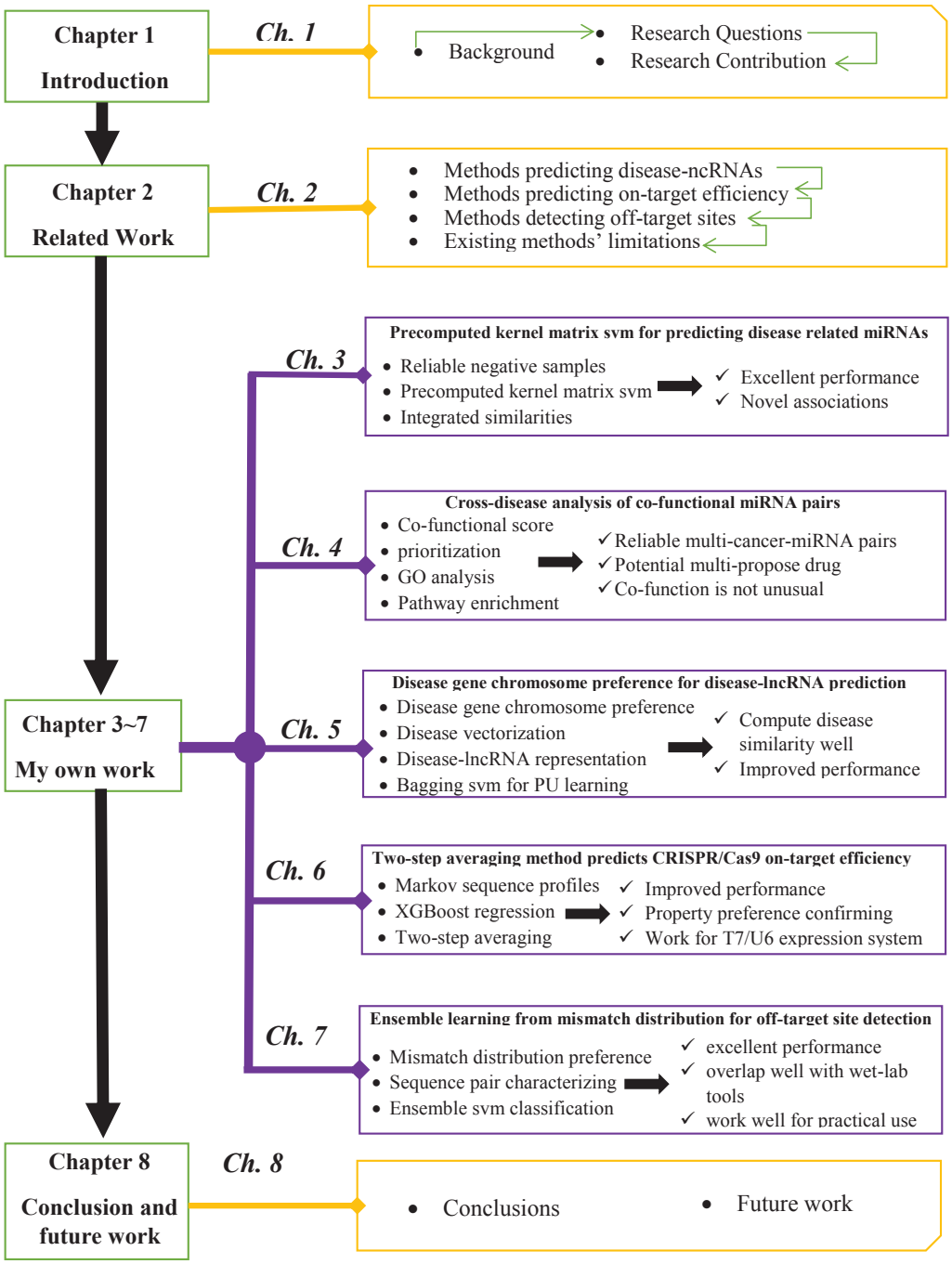


Figure 1.2: **Thesis Structure.** It includes four main parts: Introduction; Related work; My own work; Conclusion and future work. The overview of the contents in each part is shown at the right side.

# Chapter 2

# Related Work and Literature Review

This chapter describes the related studies of the work in this thesis. In Section 2.1, the existing disease-ncRNA association prediction methods are surveyed. Then, in the following Section 2.2, the strategies for CRISPR/Cas9 system on-target cutting efficiency prediction are reviewed. Later, Section 2.3 presents the published methods for detecting off-target sites of the CRISPR/Cas9 system. In Section 2.4, the limitations of these existing methods for solving disease-ncRNA gene prediction and disease-gene editing optimization problems are discussed. At last, Section 2.5 summaries the content in this chapter.

## 2.1 Disease-ncRNA Association Prediction

As the disease-miRNA association prediction and disease-lncRNA association prediction are similar problems, this section describes their related work together. We classify those existing prediction methods into four main types such as network analysis methods, similarities analysis or semi-supervised methods, supervised learning methods and other kinds of methods. Most of these methods predicted the candidate disease-ncRNA associations with the

same criterion: the similar diseases always associate with similar ncRNAs. In the following subsections, some representative methods of the four types are introduced.

### 2.1.1 Network analysis methods

The network analysis methods always find new associations from the already known relationships of the nodes in those disease-miRNA related networks. For example, the method proposed by Jiang et al. (Jiang, Hao, Wang, Juan, Zhang, Teng, Liu & Wang 2010) prioritized disease related miRNAs with the analysis of a human phenome-microRNAome network. It was extended from the disease-related protein-coding gene prediction method (Linghu, Snitkin, Hu, Xia & DeLisi 2009). The core idea is scoring the disease-miRNA pairs in the human phenome-microRNAome network with the hypergeometric distribution. This network was constructed with the disease phenotype similarities and miRNA function similarities. The miRNA functional similarity was computed according to the statistical analysis of the shared targets of two miRNAs and the phenotype similarity was obtained with the MimMiner (Van Driel, Bruggeman, Vriend, Brunner & Leunissen 2006). The authors claimed that their method can predict related miRNAs of a given disease accurately even without known miRNAs of this disease. Obviously, the accuracy of this method is limited by the accuracy of the similarities of the disease phenotypes and miRNA targets. As was concluded by the authors, their method used the overlapping of the miRNA targets to analyze the function similarities, which may ignore the function relationships of those targets themselves. Moreover, they computed the local similarities among the miRNAs (just considered the neighbor nodes of a given node), which may lose the global information of the whole network and finally affect the prediction accuracy.

Then, Chen et al. (Chen, Liu & Yan 2012) developed a network analysis method namely the RWRMDA. This method constructed two networks such as the miRNA-disease associations network (N1) and miRNA-miRNA

functional similarity network (N2). There are no edges between different diseases. To predict the disease associated miRNAs, those miRNAs in the network N1 were used as the seed RNAs and the random walk with restart (RWR) algorithm was implemented on network N2. Each of the miRNAs in N2 then obtained a score for indicating its possibility to associate with a disease. Different from the previous method, Chen et al. computed the global network similarities between the miRNAs. This method did not adopt the disease similarity information. Even though the prediction performance of this method on the benchmark dataset was claimed to be better than the previous method, it still relies on the similarities between miRNAs. The network structure may also influence the prediction. Obviously, if a disease has no connected miRNA in the studied network, no potential related miRNAs could be found for it.

In 2013, Shi et al. (Shi, Xu, Zhang, Xu, Li, Wang, Zhao, Jiang, Guo & Li 2013) predicted disease related miRNAs by analyzing the protein-protein interaction network (PPI) where some nodes of it are miRNA targets and disease genes. Their hypothesis for predicting disease-miRNA associations is that a miRNA may associate with a disease if they share the same related genes. Similar to RWRMDA, the RWR algorithm was adopted to evaluate the probability of a miRNA associating with a given disease. The miRNA targets and the disease genes of a special disease-miRNA pair were regarded as the seeds for RWR respectively. In this method, no similarity information was required. The PPI network provides the functional relationships of the miRNA target genes and the disease genes. This method's performance heavily depends on the structure of their PPI network and the accuracy of known miRNA-target and disease-gene information.

Recently, Xuan et al. (Xuan, Han, Guo, Li, Li, Zhong, Zhang & Ding 2015) adopted the RWR to analyze the miRNA functional similarity network for predicting disease-miRNA associations. This method is also similar to RWRMDA. The functional similarities of miRNAs were computed with their related diseases' semantic similarities but not their targets'

information. Two one-step transition matrices were constructed to guide the walker according to the calculated similarities of miRNAs and the known associations between miRNAs and diseases. To predict the disease associated miRNAs without known related miRNA for this disease, the author also built a bilayer network. It was extended from the former miRNA functional similarity network by connecting the diseases with their semantic similarity and connecting the diseases and miRNAs with known associations. By comparing this method with other existing methods such as the above Jiang's method and Chen's method, the author claimed that the involving of prior information and topology properties of the bilayer network can improve the prediction performance. However, both the functional similarities of miRNAs and the similarities between diseases were computed with the disease semantic similarities. This may lose some important information such as the miRNA target relationships and the disease-related genes' information, which affects the prediction accuracy.

Similarly, some network analysis methods were proposed for predicting disease-lncRNAs. LncRNAs are newer ncRNAs comparing to miRNAs. Thus, disease-lncRNA association prediction methods are mainly inspired by those miRNA related ones. For example, in 2014, three network analysis methods were reported such as Yang et al.'s method (Yang, Gao, Guo, Shi, Wu, Song & Wang 2014), Sun et al.'s RWRlncD (Sun, Shi, Wang, Zhang, Liu, Wang, He, Hao, Liu & Zhou 2014) and the ncPred proposed by Salvatore et al. (Alaimo, Giugno & Pulvirenti 2014). In Yang et al.'s method, a bipartite network containing the genes (coding or non-coding genes) and diseases was constructed with known lncRNA-disease associations and disease-genes. Then, the propagation algorithm was implemented on the bipartite network to obtain a score for indicating the probability of a lncRNA associating with the given disease. Although this method needs not to compute the disease similarities and lncRNA similarities, it significantly relies on the network structure. It's impossible to predict a given disease related lncRNAs if it has no already known associated lncRNA. Several

months later, Sun et al.'s RWRlncD was published. It adopted the RWR to analyze the lncRNA functional similarity network. This method is almost the same as Chen's RWRMDA. NcPred was proposed by adopting the concept of resources transfer to analyze the tripartite network that is composed of ncRNAs, RNA targets and diseases. This method requires no similarity or other biological information of the network nodes. As the testing dataset is small, whether this method can work well in bigger sample space should be further investigated.

In 2015, two similar methods using the RWR algorithm to analyze the lncRNA-disease heterogeneous network were proposed by Ganegoda et al. (Ganegoda, Li, Wang & Feng 2015) and Zhou et al. (Zhou, Wang, Li, Hao, Wang, Shi, Han, Zhou & Sun 2015) respectively. Both of these two methods constructed the networks with the disease similarities, lncRNA-lncRNA relationships and known disease-lncRNA associations. The differences are their distinct ways to weight the similarities and assigning the transition probabilities during the random walk. Ganegoda et al. applied the disease phenotype similarities and lincRNA tissue-specific similarities, while Zhou et al. adopted the lncRNA-miRNA interaction information and the disease semantic similarities. These two methods also have the same limitations comparing with above network analysis methods. Chen et al. (Chen 2015) then presented the KATZLDA to predict the associations between lncRNAs and diseases. It mainly computed the similarity of two nodes in the heterogeneous network to measure the probability of existing a link between them. The heterogeneous network integrates the lncRNA-disease association network, the lncRNA similarity network and the disease similarity network. In this method, similarities of lncRNAs were calculated by their functional similarities and expression profiles. The disease similarities were composed of disease semantic similarities and Gaussian interaction profile kernel similarities. The informative heterogeneous network ensures the KATZLDA can predict the disease related lncRNAs at large scale even without known associated lncRNAs for a given disease. However, the requirement of

abundant information makes the collection of those information a heavy job (or sometimes impossible).

### 2.1.2 Similarities or semi-supervised methods

Those semi-supervised or similarity based methods are also commonly used to predict the ncRNA-disease associations. This kind of methods do not depend on the network topology properties. The ncRNAs' similarities, diseases' similarities and the known associations are the main prior information. These methods regard the association prediction problem as an optimization question.

For example, the HDMP method designed by Xuan et al. (Xuan, Han, Guo, Guo, Li, Ding, Liu, Dai, Li, Teng et al. 2013) applied the idea of considering a known disease-miRNA's top-$k$ most similar neighbors as new disease-miRNAs. The miRNAs' functional similarities were computed with their related diseases' semantic, term and phenotype similarities. Each disease-miRNA pair was assigned a score to indicate whether the miRNA in this pair correlates with that disease. During the scoring, for a given disease's candidate miRNAs, their miRNA family and cluster information were also considered. The parameter $k$ was tuned by the comparison of known associations' scores with those of the unknown disease-miRNA pairs. This method is sensitive to the accuracy of miRNAs' similarities. At the same time, if a disease has no already known related miRNAs, the prediction is impossible. Furthermore, selection of the parameter $k$ is quite difficult for a large scale of diseases. In the same year, Chen et al. (Chen & Zhang 2013) proposed three models such as MBSI, PBSI and NetCBI. MBSI applied the miRNA-miRNA functional similarity for predicting disease-miRNA associations. PBSI used the disease phenotype similarities and NetCBI adopted the disease-miRNA network consistency information. These three models also leveraged the similar criterion as previous HDMP. Thus, they have the same limitations comparing to the HDMP.

Chen et al. (Chen & Yan 2014) proposed their RLSMDA, a representative

semi-supervised method, to predict disease-miRNA associations. A regularized least square function was designed to optimize the classification functions of the miRNA space and disease space. The two classification functions were then combined to output the final prediction scores indicating the possibilities of miRNAs associating with the diseases. This RLSMDA method has three kinds of advantages: negative samples are not required; it can predict different diseases' related miRNAs simultaneously; it runs fast. However, RLSMDA also has some drawbacks. First of all, it heavily depends on the miRNAs' similarities and diseases' similarities. Secondly, the model training and evaluation also involved unlabeled disease-miRNA pairs, which may introduce bias. In fact, some of those unknown pairs may be positive. At last, as were mentioned by the authors, they did not solve the parameter selection problem.

In 2013, Chen et al. (Chen & Yan 2013) presented their LRLSLDA method to predict the relationships between disease and lncRNAs. This algorithm requires to compute the diseases' similarities (Gaussian interaction profile kernel similarity) and lncRNAs' similarities (Gaussian interaction profile kernel similarity and expression similarity) as well. Then, the laplacian regularized least squares framework was adopted to predict the associations. LRLSLDA is also a semi-supervised prediction method, thus it may have the similar drawbacks as the RLSMDA, e.g. similarity dependency, unlabeled samples related bias.

### 2.1.3 Supervised learning methods

To date, some supervised learning methods were adopted to address the problem of predicting disease-ncRNA associations. The first supervised learning method may be the one proposed by Xu et al. (Xu, Li, Lv, Li, Xiao, Shao, Huo, Li, Zou, Han et al. 2011). In this method, the authors represented miRNAs with topological features that were extracted from the miRNA-target network. They used the known disease-miRNAs as positive samples. Their negative samples were selected according to the tissue-

specificity properties of the miRNAs. A SVM classifier was built to conduct binary classification of the miRNAs for a given disease. Based on their case study on prostate cancer, the author drew the conclusion that their network-centric method can be adopted to prioritize novel disease-miRNAs. They conducted the *in vitro* experiment to verify the reliability of their prediction results. There are several disadvantages of this method. Firstly, only the network topological features were adopted for the prediction. Other useful features can be included such as the miRNA sequence and function. Secondly, the negative samples are not reliable enough as the miRNA with low expression level is not equal to it will not dys-express. Furthermore, this method is not suitable for large scale prediction. For each disease, a specific classifier needs to be constructed. At last, their sample size is quite small, there are just 37 positive samples and 44 negative samples. This may lead to inadequate learning of the machine.

Another SVM classification method was published by Jiang et al. (Jiang, Wang, Jin, Li & Wang 2013). In Jiang's method, a miRNA is represented by a vector where the similarities between this miRNA and other miRNAs are the elements. Similarly, a disease is vectorized by computing the similarities of it comparing with other diseases. The two vectors were combined to characterize the disease-miRNA pair and used as the input of the SVM to train the prediction model. This method's negative samples were selected randomly from those unknown disease-miRNA pairs. Apparently, this method has the following three defects. Firstly, the feature vector may change if the members of miRNAs or diseases change. Then, those negative samples are not reliable, which may result in a high false positive rate. At last, the sample size of this method is also small, only 270 positive samples are used in the cross-validation.

In 2015, Zhao et al. (Zhao, Xu, Liu, Bai, Xu, Xiao, Li & Zhang 2015) proposed a naive bayesian classifier to find potential cancer-related lncRNAs with genome, regulome and transcriptome information. This method also regarded the random selected unknown disease-lncRNA pairs as negative

samples. Their feature space is complex as much information were required.

### 2.1.4 Other types of prediction methods

Apart from the previous three types of methods, some other strategies were applied to solve the disease-ncRNA association prediction problems. In the year 2014, Biswas et al. (Biswas, Gao, Zhang & Wu 2014) adopted the idea of non-negative matrix factorization (NMF) to predict the disease related lncRNAs. At first, a relationship matrix was built with the known disease-lncRNA associations. Then, different types of NMF models were applied to obtain special vectors for each lncRNA and each disease. Finally, the scores of the lncRNA-disease pairs were calculated according to their corresponding vectors before ranking all the unknown pairs. This method did not apply the similarities, network and negative samples. It seems easy to be implemented as little prior information was required. However, this may also reduce the accuracy of the prediction results.

Recently, Wang et al. (Wang, Ma, Ma, Chen, Yang, Xi & Cui 2016) developed a sequence-based bioinformatics tool for identifying disease related lncRNAs. This method predicted the miRNAs that interact with the given lncRNA at first. Then, these miRNAs were enriched to known disease-associated miRNA sets. Those diseases whose miRNA sets are significantly enriched were thought to be associated with the given lncRNA. This method applied the idea of those disease-miRNAs regulated lncRNAs are more likely to associate with the diseases. Two types of information are critically important such as the lncRNA-miRNA interactions and the disease-miRNA relationships, which affect the method's performance a lot.

A newer miRNA-disease association prediction method was proposed by Pasquier et al. (Pasquier & Gardès 2016) later. This method firstly represented diseases and miRNAs with high-dimensional vectors according to known distributional information such as miRNA-neighbor associations, miRNA-family associations, known miRNA-disease associations and so on. After dimensionality reduction, the similarity between a disease vector and a

miRNA vector was used to evaluate the possibility of this miRNA associating with the disease. The involved information of this method is abundant, which improves the prediction performance. However, to extract all those required information is difficult at the same time.

## 2.2 CRISPR/Cas9 On-target Cutting Efficiency Prediction

The CRISPR/Cas9 system is one of the most promising tools to edit those disease genes (including ncRNA genes) currently (Ho, Zhou, Huang, Koirala, Xu, Fung, Wu & Mo 2014, Swiech et al. 2015, Wu, Zhou, Fan, Zhang, Zhang, Wang, Xie, Bai, Yin, Liang et al. 2015, Zhu, Li, Liu, Chen, Liao, Xu, Xu, Xiao, Cao, Peng et al. 2016). Selecting an optimal spacer sequence to make sure that the system has higher cutting efficiency is an important step for CRISPR/Cas9 system design. Evaluating the cutting efficiencies by biological experiments is accurate but expensive and time-consuming. Computational methods are always fast, cheap and easy, which can be used to select those possible optimal spacers for further experimentally confirmation. They can help the CRISPR/Cas9 system design by saving time and cost in practical use. This section reviews the existing computational methods for CRISPR/Cas9 system on-target cutting efficiency prediction. The known methods are classified into three categories such as classification, regression and non-machine learning methods. They are described in three subsections below. The basic descriptions of these methods are shown in following Table 2.1.

### 2.2.1 Binary classification of the sgRNAs

The binary classification methods always assign a given sgRNA the label of highly-active or low-active. In 2014, Doench et al. (Doench, Hartenian, Graham, Tothova, Hegde, Smith, Sullender, Ebert, Xavier & Root 2014)

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

| Tool | method | year | server | offline | author |
|---|---|---|---|---|---|
| Rule set 1 | logistic regression classifier | 2014 | yes | no | Doench *et al.* |
| SSC | Elastic-Net model classification | 2015 | yes | yes | Xu *et al.* |
| sgRNAscorer1.0 | SVM classifier | 2015 | yes | yes | Chari *et al.* |
| WU-CRISPR | SVM classifier | 2015 | yes | yes | Wong *et al.* |
| CRISPRScan | liner regression | 2015 | yes | no | Moreno-Mateos *et al.* |
| BiophyM | Biophysical Model | 2016 | no | no | Farasat *et al.* |
| ge-CRISPR | SVM classification/regression | 2016 | yes | no | Kaur *et al.* |
| Rule set 2 | boosted regression tree | 2016 | yes | yes | Doench *et al.* |
| sgRNAscorer2.0 | SVM classifier | 2017 | yes | no | Chari *et al.* |
| DeepCRISPR | deep learning | 2018 | yes | yes | Chuai *et al.* |
| bacteriaSgRNA | boosting regression tree | 2018 | no | yes | Guo *et al.* |

reported their classification method (Rule set 1) for helping select highly active sgRNAs. The cutting efficiencies of 1841 sgRNAs for editing 9 genes were measured with wet lab experiments. For each gene's sgRNAs, their efficiencies were normalized to be scores between 0 to 1. After ranking all the sgRNAs' efficiency scores, the top-ranked 20% sgRNAs were regarded as highly-active ones. The sgRNAs were characterized with features including individual nucleotides indexed by position in the 30-mer target site ($N_4N_{20}NGGN_3$, where $N_{20}$ is the spacer sequence and $NGG$ is the PAM), all pairs of adjacent nucleotides indexed by position in the 30-mer target site and the number of 'G' and 'C' in the 20nt spacer. Then, an L1-regularized linear support vector machine was adopted to select important features from the feature space with the L1-norm penalty. At last, a logistic regression classifier was trained with these selected features. According to the authors' analysis, they confirmed that there exists nucleotide preference of the highly active sgRNAs such as preferring guanine ('G') but disliking cytosine ('C') at the 20th position (counting from 5' to 3'), preference for cytosine and against guanine at position 16. The dataset used by this method has been one of the golden standard datasets in sgRNA efficiency classification.

In the year 2015, three another sgRNA efficiency classifiers were published by different research groups. Xu et al. (Xu, Xiao, Chen, Li, Meyer, Wu,

Wu, Cong, Zhang, Liu et al. 2015) built an Elastic-Net model (SSC) for the classification of the sgRNAs with only sequence features such as the binary vector for representing the presence or absence of the nucleotides. They collected 3 datasets containing the sgRNAs for editing human and mouse essential genes to train and evaluate their method. The authors also confirmed some similar nucleotide preferences as Doench et al.'s work (Doench et al. 2014). In addition, Xu et al. observed new preferences such as the cytosine preference at the 3rd position upstream to the PAM. By testing with the independent datasets, the authors drew the conclusion that their model performs better than Doench et al.'s method.

Later, Chari et al. (Chari, Mali, Moosburner & Church 2015) proposed their in vivo library-on-library method to simultaneously measure the sgRNAs' activities (sgRNAscorer1.0). In Chari et al.'s work, they collected two relatively small datasets for $Cas9_{Sp}$ (from Streptococcus pyogenes) and $Cas9_{St1}$ (from Streptococcus thermophilus, with the PAM of 'NNAGAAW') respectively. The spacer+PAM sequences were encoded with a 4-bit binary system for each nucleotide as the inputs of the SVM to construct their classifier. Comparing with Doench et al.'s method, the authors observed a modest correlation between the two methods' results. This was explained by the reasons that these two classifiers' training dataset are different and their experiment time is not the same. In addition, Chari et al. applied the direct sequence-based readout to assay the sgRNA activity while Doench et al. used a phenotype-based readout. Chari et al. also confirmed similar nucleotide preferences of highly active sgRNAs and they found that these preferences are generally existing (both for $Cas9_{Sp}$ and $Cas9_{St1}$).

Wong et al. (Wong, Liu & Wang 2015) reported their SVM classifier for the sgRNA cutting efficiency classification (WU-CRISPR). In their work, CRISPR related RNA-seq data were analyzed. This helped them identify many novel features to characterize sgRNAs for the classification. Those novel features include the repetitive bases, overall nucleotide usage and structural characteristics such as overall secondary structure, self-folding

free energy, and the accessibility of individual nucleotides in the structure. By comparing with the above three existing methods, the authors claimed that their classifier works better than the other methods according to the Precision-Recall (PR) curves.

In the year 2016 and 2017, some other methods for classifying sgRNAs were proposed. For example, Kaur et al. (Kaur, Gupta, Rajput & Kumar 2016) reported their ge-CRISPR which integrates many different features and applies feature selection to improve their classifier's performance. Chari et al. (Chari, Yeo, Chavez & Church 2017) provided an improved version tool sgRNAscorer2.0 which can predict sgRNA activities across multiple CRISPR systems.

### 2.2.2 Regression methods for sgRNA cutting efficiency prediction

The regression methods prone to give each sgRNA a continuous score for indicating its cutting efficiency. For instance, Moreno-Mateos et al. (Moreno-Mateos, Vejnar, Beaudoin, Fernandez, Mis, Khokha & Giraldez 2015) built one of the earliest regression methods namely CRISPRScan to predict the sgRNAs' cutting efficiencies. They firstly measured 1280 sgRNAs' efficiencies for targeting 128 zebrafish genes and normalized them to the scores between 0 to 100 (100 means highest). After the filtering process, 1020 sgRNAs with experimentally determined efficiencies were remained for building their linear regression model. The selected subset (91-dimensional) of the mononucleotide type indexing (4 bits) and dinucleotide type indexing features were applied to represent the 35nt target site sequences ($N_6N_{20}NGGN_6$, NGG is the PAM and $N_{20}$ is the spacer). The cross-validation and independent tests proved that their predicted efficiency scores correlate well with the experimentally measured values. In addition, they have found some determinants for the sgRNA's cutting efficiency. For instance, the guanine enrichment and adenine depletion can increase the sgRNA stability and activity.

In 2016, Doench et al. (Doench et al. 2016) improved their previous work (Rule set 1 (Doench et al. 2014)) by proposing the Rule set 2 for the prediction of sgRNA cutting efficiency. They expanded their dataset with another 2549 sgRNAs whose efficiencies were determined with the drug resistance phenotype. The total 4379 sgRNAs (some sgRNAs were filtered out as they were mapped to multi-genome locus) with their normalized efficiency scores (0 to 1) form one of the most popular gold standard datasets. By combining their previous sequence features with newly proposed ones such as the melting temperature and the sgRNA location within the protein properties, the gradient-boosted regression trees were trained to build their regression model. Through the performance comparisons, the authors concluded that their Rule set 2 outperforms their Rule set 1 (Doench et al. 2014), Xu et al.'s SSC (Xu et al. 2015) and Wong et al.'s (Wong et al. 2015) WU-CRISPR.

Recently, two more regression methods for sgRNA cutting efficiency prediction were published. The first one is the DeepCRISPR proposed by Chuai et al. (Chuai, Ma, Yan, Chen, Hong, Xue, Zhou, Zhu, Chen, Duan et al. 2018), which applied the deep learning to predict the efficiency scores by learning the sequence information and the epigenetic information of the target sites. They tested their model and others' with various testing scenarios to prove their method's excellent performance. Later, Guo et al. (Guo, Wang, Guan, Liu, Luo, Xie, Zhang & Xing 2018) reported their gradient boosting regression tree-based regression model (bacteria sgRNA). In their method, the sgRNAs were encoded by the strategy similar to Doench et al.'s Rule set 1 (Doench et al. 2014). Guo et al.'s method is mainly designed for bacteria but not mammals. They proved that the methods designed for mammals may cannot predict the sgRNAs' cutting efficiencies well for bacteria.

### 2.2.3 A Non-machine learning method for on-target cutting efficiency prediction

Though machine learning methods are popular for the CRISPR/Cas9 on-target site cutting efficiency prediction, there exists a biophysical model method for completing this task. Farasat et al. (Farasat & Salis 2016) developed free energy models to estimate the cutting efficiency of the sgRNA. It adopted the statistical thermodynamics and kinetics to model the processes of Cas9:crRNA complex formation, diffusion, site selection, reversible R-loop formation and cleavage. Then, abundant structural, biochemical, expression, and next generation sequencing data were used to determine their model's parameters.

## 2.3 CRISPR/Cas9 Off-target Site Detection

Though CRISPR/Cas9 system can generate DNA cleavage with high efficiency, the off-target effect is one of the important limitations for its practical usage. Detecting the potential off-target sites is a significant step for the optimal design of a sgRNA. This section surveys the existing methods for detecting genome-wide off-target sites of a given sgRNA. These methods can be classified into wet-lab technologies and computational methods. For the computational methods, mismatch information scoring methods and machine learning methods are mainly included. These three types of methods are discussed in the following three subsections. The basic information of them are listed in following Table 2.2.

### 2.3.1 Wet-lab technologies for off-target site detection

Some wet-lab technologies have been proposed to detect off-target sites of the CRISPR/Cas9 system. These technologies apply the next-generation sequencing (NGS) technology to detect mutated reads that are induced by the CRISPR/Cas9 system. These wet-lab tools can detect bona fide off-

Table 2.2: **The existing off-target site detection methods.**

| Tool | method | year | web-server | offline tool | author |
|---|---|---|---|---|---|
| CIRCLE-seq | wet-lab | 2017 | no | yes | Tsai *et al.* |
| Digenome-seq | wet-lab | 2015 | no | yes | Kim *et al.* |
| multiplex Digenome-seq | wet-lab | 2016 | no | res | Kim *et al.* |
| GUIDE-seq | wet-lab | 2014 | no | yes | Tsai *et al.* |
| HTGTS | wet-lab | 2014 | no | yes | Frock *et al.* |
| SITE-Seq | wet-lab | 2017 | no | yes | Cameron *et al.* |
| CCTop | scoring | 2015 | yes | yes | Stemmer *et al.* |
| CFD | scoring | 2016 | yes | yes | Doench *et al.* |
| CRISPOR | integration | 2016 | yes | yes | Haeussler *et al.* |
| CRISTA | RandomForest | 2017 | yes | yes | Abadi *et al.* |
| CROP-IT | scoring | 2015 | yes | no | Singh *et al.* |
| MIT-score | scoring | 2013 | yes | no | Hsu *et al.* |
| Elevation | machine learning | 2018 | yes | no | Listgarten *et al.* |
| DeepCRISPR | deep learning | 2018 | yes | yes | Chuai *et al.* |

target sites at genome scale.

The first widely used wet-lab technology is the GUIDE-seq (Tsai, Zheng, Nguyen, Liebers, Topkar, Thapar, Wyvekens, Khayter, Iafrate, Le et al. 2015). It can detect DSBs caused by CRISPR-Cas nucleases such as the CRISPR/Cas9. This technology detects the off-target site by two procedures. Firstly, DSBs in the genomes of living human cells are tagged by integration of a blunt, double-stranded oligodeoxynucleotide (dsODN) at these breaks by means of an end-joining process consistent with NHEJ. Then, dsODN integration sites are detected with the NGS technology. Through the validation with previous datasets and their low-throughput experiments, the authors reported that GUIDE-seq can detect genome-wide unbiased off-target sites accurately.

Frock et al. (Frock, Hu, Meyers, Ho, Kii & Alt 2015) applied their previously proposed technology high-throughput, genome-wide translocation sequencing (HTGTS) (Chiarle, Zhang, Frock, Lewis, Molinie, Ho, Myers, Choi, Compagno, Malkin et al. 2011) to detect CRISPR/Cas9 off-target sites. HTGTS used the linear amplification-mediated PCR (LAM-PCR) to distinguish genome-wide DSBs generated by engineered nucleases (e.g. CRISPR/Cas9) from those endogenous or ectopic ones. HTGTS was claimed

to have the energy of identifying nuclease-generated, on-target and off-target DSBs and associated collateral chromosomal damage.

Kim et al. (Kim, Bae, Park, Kim, Kim, Yu, Hwang, Kim & Kim 2015) designed their tool Digenome-seq for profiling the genome-wide off-target sites of the CRISPR/Cas9 system. The Digenome-seq adopts the whole genome sequencing to identify off-target mutations via sequencing in vitro nuclease-digested genomes. The differences between digested genome reads and the traditional reads are that the former reads are vertically aligned at cleavage sites while the later reads would be aligned in a staggered manner. By comparing with the GUIDE-seq and HTGTS, the authors concluded that the three tools obtain comparable results but their tool has the advantage of without chromatin accessibility limitation. Kim et al. (Kim, Kim, Kim, Park & Kim 2016) improved their Digenome-seq by developing the multiplex Digenome-seq. This new tool can profile up to 11 CRISPR/Cas9 systems' genome-wide specificities simultaneously to help save time and reduce cost. They also proved that the multiplex Digenome-seq detects more complete bona fide off-target sites than the other wet-lab technologies.

In 2017, the tool SITE-Seq was presented by Cameron et al. (Cameron, Fuller, Donohoue, Jones, Thompson, Carter, Gradia, Vidal, Garner, Slorach et al. 2017). This tool detects the off-target sites by assaying mutations at each cut site using amplicon sequencing. The authors designed this tool for resolving the issues existing in the GUIDE-seq and HTGTS that they rely on cellular events such as the integration of donor sequences or chromosomal translocations. In addition, comparing to the Digenome-seq, it requires no high read depth. The authors found that their SITE-Seq can detect all the GUIDE-seq, HTGTS and Digenome-seq detected off-target sites and also some new sites, which proves their method's excellent performance.

At the same time, Tsai et al. (Tsai, Nguyen, Malagon-Lopez, Topkar, Aryee & Joung 2017) reported their tool CIRCLE-seq for in vitro detection of CRISPR/Cas9 off-target sites. Their core idea is selectively sequencing the Cas9-cleaved genomic DNA. During CIRCLE-seq's work, the genomic DNA

is firstly sheared and circularized by intramolecular ligation. Then, only the circular DNA molecules containing a Cas9 cleavage site can subsequently be linearized and sequenced with NGS. The authors concluded that their tool is accessible, rapid and comprehensive. It applies a highly sensitive, sequencing-efficient in vitro screening strategy, which outperforms the other cell-based tools (e.g., GUIDE-seq, HTGTS). It was also reported to have the advantage of identifying off-target mutations associated with cell-type-specific single-nucleotide polymorphisms.

### 2.3.2 Mismatch information scoring methods for off-target site detection

The mismatch number and mismatch type between the sgRNA sequence and its target region sequence mainly determine whether there is an off-target cutting. Detecting off-target sites by directly observing the mismatch information is a popular strategy. We call those computational tools applying this strategy the mismatch information scoring methods.

Hsu et al. (Hsu, Scott, Weinstein, Ran, Konermann, Agarwala, Li, Fine, Wu, Shalem et al. 2013) proposed their tool for evaluating sgRNA specificity by a scoring method (MIT-score). At first, they observed the phenomenon that a CRISPR/Cas9 system tolerates mismatches at different positions in a sequence-dependent manner. The off-target cutting is sensitive to the number, position and distribution of mismatches between the sgRNA and its target. These were found by measuring the cutting efficiencies of the sgRNA variants containing mismatches with different mismatch number, at various positions and with diverse mismatch types. Finally, they constructed a scoring function to compute a score for indicating the off-target cutting efficiency at a given site (Their detail scoring function can be found from the website: http://crispr.mit.edu/about).

Stemmer et al. (Stemmer, Thumberger, del Sol Keyer, Wittbrodt & Mateo 2015) developed the CCTOP to detect off-target sites of the sgRNAs. This method firstly searches all the potential target sites of a given sgRNA

with Bowtie (Langmead, Trapnell, Pop & Salzberg 2009), a popular read aligner. Then a score for each potential off-target site can be computed according to its mismatch number and positions. The simple scoring function is shown as formula 2.1, where *pos* is the position of a mismatch, counted from the 5' end and the base 1.2 is experimentally determined. This score was used to indicate the likelihood of a stable sgRNA/DNA heteroduplex, which also can be regarded as an indicator of the possibility of existing an off-target cutting.

$$score_{off-target} = \sum_{mismach} 1.2^{pos} \tag{2.1}$$

Singh et al. (Singh, Kuscu, Quinlan, Qi & Adli 2015) designed their algorithm CROP-IT to predict CRISPR/Cas9 off-target sites. The CROP-IT considers both the sequence information and the whole genome level biological information to compute the possibility of a site to be an off-target site. It filtered the potential off-target sites with the requirement of the PAM to be 'NGG' or 'NNG' at first. Then, the score of a given site was computed by combining the weighted segmented sequence score and the chromatin state score. The detail scoring function can be found from their published paper. The authors claimed that their method outperforms the existing computational methods such as the above MIT-score (Hsu et al. 2013) and CCTOP (Stemmer et al. 2015).

In 2016, Doench et al. (Doench et al. 2016) proposed their tool CFD score to measure the off-target cutting efficiency. This tool was designed by firstly investigating the activities of the sgRNAs when insertion, deletion or mismatches were introduced into to them. Then, they examined the changes in activity produced by the three different types of variants. At last, their CFD score function was produced according to their collected data, where only the mismatch type and position information were considered. Haeussler et al. (Haeussler, Schönig, Eckert, Eschstruth, Mianné, Renaud, Schneider-Maunoury, Shkumatava, Teboul, Kent et al. 2016) evaluated previously

mentioned scoring methods with their collected datasets. They found that CFD score achieves better performance than the other tools.

### 2.3.3 Machine learning methods for off-target site detection

Recently, three machine learning methods were published for solving the sgRNA specificity evaluation issue. The first one is the CRISTA of Abadi et al. (Abadi, Yan, Amar & Mayrose 2017). CRISTA collected validated off-target cutting activity data from those published datasets and assembled selected uncleaved sites whose activities are 0. These collected sites were characterized with sequence features such as PAM sequence, nucleotide composition and GC content, chromatin structure, sgRNA secondary structure, similarities between sgRNAs and those sites and the mismatch or bulge information. Then, a random forest regression model was constructed to predict the off-target cutting activity of a given site. By selecting an appropriate threshold, it also can label a site as cleaved or uncleaved. Comparing with those scoring methods, the authors confirmed that CRISTA achieves better performances under different tests.

Listgarten et al. (Listgarten, Weinstein, Kleinstiver, Sousa, Joung, Crawford, Gao, Hoang, Elibol, Doench et al. 2018) created their cloud-based service tool Elevation to predict off-target activities. The tool assigned scores to individual guide-target pairs at first and also aggregates them into a single, overall summary guide score. The authors also concluded that their Elevation outperforms the existing tools.

Most recently, Chuai et al. (Chuai et al. 2018) proposed their tool DeepCRISPR which contains the function of evaluating sgRNA target specificity to help design sgRNAs. At first, this method fitted a given sgRNA and its one off-target site sequences into the pre-trained DCDNN-based network respectively for sample representation. Then, the two output networks were combined channel-wisely to train their CNN classifier for classifying a sgRNA-potential site pair into positive or negative. The authors

applied various testings to prove their tool's good performance comparing to the scoring methods.

## 2.4 Limitations of Existing Methods

This section concludes the limitations of those existing methods for disease-ncRNA association prediction and the optimal design of a CRISPR/Cas9 system. They are discussed in two separate subsections corresponding to the two research topics.

### 2.4.1 Limitations of the disease-ncRNA association prediction methods

Previously, we reviewed four types of existing methods for predicting disease-ncRNA associations. Each of them has some limitations. For example, the network analysis methods have the main limitation of strongly relying on the network structures' completeness. For those complex networks, vast prior information are required. This not only introduces more incompleteness to the network but also makes the mining of new connections extremely difficult.

Comparing with the network analysis methods, the similarity or semi-supervised methods have no network structure dependency. However, this type of methods always need to compute accurate disease similarities and ncRNA similarities. Sometimes, selecting suitable thresholds for similarity analysis methods or tuning the parameters of the semi-supervised methods are not easy. In addition, the model optimization and performance evaluation processes adopted the unknown relationships, which may introduce bias to the final prediction models.

Those supervised learning methods are rare to date. The biggest limitation is that reliable negative samples are required. The existing methods' strategies to collect negative samples are not reliable enough. Secondly, bigger datasets should be collected to optimize and evaluate

the prediction models. In addition, constructing effective feature space to characterize the disease-ncRNA pairs is also a difficult problem to be solved.

At last, the forth type of methods are different from the other three types as they borrow ideas from other fields. These methods avoid the limitations of those traditional methods. However, they are always complex and are not suitable for large-scale prediction.

### 2.4.2 Limitations of the methods for gene editing optimization

Three types of computational methods have been published to predict the CRISPR/Cas9 on-target cutting efficiency such as classification methods, regression methods and the biophysical model method. Each of them has some limitations that should be addressed. Those classification methods have two kinds of limitations. Firstly, the definitions of highly-active and low-active sgRNAs are not consistent and objective. For example, in Doench et al.'s Rule set 1 (Doench et al. 2014), those top-ranked 20% sgRNAs were defined as highly-active while the remaining ones were labeled as low-active. However, in Wong et al.'s work (Wong et al. 2015), the top 20% and bottom 20% were regarded as high and low respectively. Some other rules were also used such as considering the overlaps of the top quartiles and bottom quartiles of those sgRNAs (Chari et al. 2015) or observing the sgRNAs' decline in abundance in the screens (Xu et al. 2015). These various definitions may affect the prediction models. Comparisons between different methods that apply inconsistent sample definitions may be unfair. Secondly, classifying sgRNAs into highly-active or low-active is not suitable for helping sgRNA design in practical usage. Usually, a lot of sgRNAs are predicted to be highly-active, suggesting the best one is difficult.

The regression methods for the on-target cutting efficiency prediction also contain some limitations. First of all, the efficiency normalization is always implemented at the gene-scale but not genome-scale. Some of the existing models were trained with the datasets involving multi-genes but were tested

on a single gene. This may introduce bias during the model training and validation. Secondly, some prediction models were just trained with limited species' datasets but were applied to other not well-investigated species. Whether the genome differences affect the cutting efficiency prediction has not yet been confirmed. In addition, as was reported by Haeussler et al. (Haeussler et al. 2016), the expression systems of the sgRNAs also make sense to the prediction performance. The sgRNAs expressed from the T7 and U6 expression systems should be predicted separately. At last, current regression methods' performances are not good enough. More work should be done to improve the prediction accuracy such as extracting more meaningful features for sgRNA representation or designing new prediction strategies.

The biophysical model method predicts the on-target cutting efficiency through the simulation of CRISPR/Cas9-DNA interaction with mathematical models. This method seems more complex than those machine learning methods. Much information are required to optimize the models' parameters. This method is not suitable for predicting the on-target cutting efficiencies at large scale.

As is described previously, there are wet-lab technologies, scoring methods and machine learning models to detect CRISPR/Cas9 off-target sites. Those wet-lab technologies can detect bona fide off-target sites as they directly observe the mutations induced by gene editing events. However, these methods are always costly and time-consuming. The mismatch information scoring methods run fast and are easy to be implemented. On the other hands, these methods give no consistent threshold to say which site is a real off-target site. In addition, these methods are too simple to accurately characterize the off-target site sequences' properties. Current machine learning methods have the main limitation of lacking enough reliable training samples. Especially, the real off-target sites account for a small part of the whole genome-widely potential editing sites, which makes the datasets extremely unbalanced. Furthermore, the false positive rates and false negative rates are high. More efforts are required to solve these limitations

for the optimal design of the sgRNAs.

## 2.5 Summary

This chapter mainly reviews the existing methods for addressing the problems of disease-ncRNA association prediction, CRISPR/Cas9 system on-target cutting efficiency prediction and off-target site detection. These methods' limitations have also been discussed. In conclusion, the collection of reliable samples are crucial for designing excellent models to solve these problems. If a supervised classification method is applied, negative samples should be prepared well at first. The second point for improving the performance is to extract effective features for the representation of those samples. Extracting domain knowledge characteristics and applying feature selection strategies may exactly benefit the performance improvement.

# Chapter 3

# Precomputed Kernel Matrix SVM Method for Predicting Disease Related miRNAs

## 3.1 Introduction

As mentioned in Chapter 1, miRNAs play significant roles in disease development. The relationships between miRNAs and diseases are still remaining largely unknown. Existing computational methods for predicting disease-related miRNAs have been reviewed in Chapter 2. The key idea in the similarity measuring criterion adopted by most of these existing methods is that: similar RNAs (functionally similar) are always associated with similar diseases (phenotypically similar, genotypically similar or semantically similar). We have concluded in Chapter 2 that these existing computational methods have some limitations such as the network structure dependency, the threshold and parameter selection, and lacking reliable negative samples.

To improve the prediction performance and overcome those limitations, we propose a new method to make predictions of disease-related miRNAs. Two new ideas are explored. One is the construction of a set of reliable

negative samples of disease-miRNA association through miRNA expression comparison between control and diseased subjects. The second idea is the use of precomputed kernel matrix for support vector machines, which can avoid characterizing the samples directly and the step to tune the parameters of the kernel functions. The area under the ROC curve (AUC) performance of our method is much superior to the literature methods on benchmarking data sets. Our case studies have demonstrated that our prediction method can also work well even such a disease is given that has no currently known disease related miRNAs.

## 3.2 Method

### 3.2.1 Data sets for the diseases, miRNAs and their related genes

Diseases and miRNAs stored at different databases may have different names or IDs. To deal with this inconsistency issue, we mapped the names of the diseases and miRNAs from all the relevant databases to the database Disease Ontology (DO) (Schriml, Arze, Nadendla, Chang, Mazaitis, Felix, Feng & Kibbe 2012) and miRBase v21.0 (Kozomara & Griffiths-Jones 2014). The Medical Subject Headings (MeSH) (Lipscomb 2000) and Comparative Toxicogenomics Database (CTD) (Davis, Murphy, Saraceni-Richards, Rosenstein, Wiegers & Mattingly 2009) were used as the dictionaries of the disease names. We searched in DO for all the disease names of a data set. When exact terms were found in DO, the names and the DO ids were recorded and stored in a separate file. Otherwise, we searched in MeSH and CTD and used their synonyms to map them to DO terms. To map the names of the miRNAs, we searched the given ids of the miRNAs in miRBase v21. When a term was not found, then it was discarded (according to miRBase, it may be a dead record because it is not a miRNA, or the record has been replaced by another one). A miRNA id is always related

to two mature miRNA ids with the suffix of '-5p' or '-3p' which means a precursor miRNA will generate two mature miRNAs from the 5'-arm or the 3'-arm respectively. As the mature miRNAs are the real functional parts, the miRNAs from different resources were mapped to the mature miRNA ids in miRBase v21. For those older version ids, we also mapped them to the current mature miRNA ids according to the term of Previous IDs of the miRBase database. Finally, each miRNA was mapped to one mature miRNA id of the database miRBase v21.

The genes were mapped to the entrez gene ids according to the HUGO Gene Nomenclature Committee (HGNC) (Povey, Lovering, Bruford, Wright, Lush & Wain 2001). To get the disease-related genes, we downloaded the supplementary files of (Cheng, Li, Ju, Peng & Wang 2014) which contains 117,190 associations between 2817 diseases and 12063 genes from the database SIDD (Cheng, Wang, Li, Zhang, Xu & Wang 2013). After data correction and redundancy removal, we obtained a data set of 114754 disease-gene associations between 2802 diseases and 10893 genes. To get the target genes of those miRNAs, we searched two databases: miRecords (Xiao, Zuo, Cai, Kang, Gao & Li 2009) and miRTarBase (Hsu, Lin, Wu, Liang, Huang, Chan, Tsai, Chen, Lee, Chiu et al. 2010). After mapping the miRNAs to miRBase v21 and mapping the gene names to entrez gene ids, we retrieved 322,269 miRNA-target pairs between 2588 miRNAs and 14794 genes. These disease genes and miRNA targets are stored in Supplementary files 1 and 2.

### 3.2.2 Positive samples and negative samples for training the prediction model to identify unknown disease-miRNA associations

There are several disease-miRNA databases such as miR2Disease (Jiang et al. 2009), HMDD (Lu, Zhang, Deng, Miao, Guo, Gao & Cui 2008), and miRCancer (Xie, Ding, Han & Wu 2013). This work focuses on the human mature miRNAs. The database HMDD stores the miRNAs as the

precursor miRNA ids, these ids were first converted into mature miRNA ids according to the provided reference links before mapping them to the mature miRNA ids. After mapping the miRNAs and diseases to miRBase v21 and DO respectively, we retrieved 4578 associations between 463 miRNAs and 263 diseases from HMDD, 1952 associations between 83 cancers and 341 miRNAs from miRCancer, and 2096 disease-miRNA associations between 108 diseases and 287 miRNAs from miR2Disease. These are known disease-miRNA associations and they are used as the positive samples for the training of the prediction model.

Selection of negative samples, i.e., those disease-miRNA pairs that have little associations, is a difficult problem. We explored a novel idea to select credible negative samples. The new idea is to select negative samples according to the expression data of the miRNAs that we downloaded from the Gene Expression Omnibus (GEO) database (Edgar, Domrachev & Lash 2002). We computed the fold changes of the miRNAs in the diseased patients comparing with the controls (i.e., the adjacent normal cells or the healthy contributor's corresponding cells) according to the given platform information of the GEO database. A disease-related miRNA is always differentially expressed significantly between these two groups of subjects. Those miRNAs that are not significant differential expressed (the fold changes smaller than 0.05) will be regarded as non-disease related miRNAs. After conducting analysis on 78 GSE accessions (some accessions without enough information for computing the fold changes were removed), we determined 21432 disease-miRNA pairs between 2473 miRNAs and 73 diseases which have little association. The accession ids can be found in the Supplementary file 3. By comparing this data set of negative samples with the above HMDD-based, miRCancer-based and the miR2Disease-based positive data sets, those pairs that appeared in both of the negative data set and the positive data sets were discarded. We then obtained 4041, 1838 and 1487 disease-miRNA pairs respectively from HMDD, miRCancer and miR2Disease, which were regarded as positive samples. 20772 disease-

miRNA pairs extracted by the analysis of the GSE accessions were used as negative samples. To obtain more reliable negative samples, we also removed those diseases that have no known related miRNAs and those miRNAs that have no known related diseases according to the three positive data sets from the 20772 disease-miRNA pairs. Finally, there are 4638 negative samples involving 53 diseases and 538 miRNAs. All these four data sets are further described in Supplementary file 4.

We note that Jiang's method (Jiang et al. 2013) takes all those unknown disease-miRNA pairs as negative samples and constructed balanced data sets by a random selection of a subset of the negative samples as the same size of the verified disease-miRNA associations. Xu's method (Xu, Li, Lv, Li, Xiao, Shao, Huo, Li, Zou, Han et al. 2011) takes those miRNAs at the lowest expression levels in the normal tissue as negative samples. Our method for selecting negative samples is different and more convincing as we consider the fold changes of the expression levels of the miRNAs between diseased and control tissues.

### 3.2.3 Precomputed kernel matrices for the support vector machine

We applied support vector machine (SVM) to predict disease-related miRNAs. SVM is a supervised learning model for classification and regression (Cortes & Vapnik 1995). We adopted the LibSVM version 3.20 (Chang & Lin 2011) in this work. Usually, one can extract the features of the samples as the input of SVM to implement classification or regression with different kernel functions such as linear kernel, polynomial kernel, radial basis function kernel. However, even though we can represent a miRNA as a feature vector, it is hard to design an appropriate feature vector to describe a disease. Diseases are always phenotypes of patients. It is difficult to find the common properties of diseases that can be normalized as mathematical variables. To overcome this issue, we proposed to use precomputed kernel matrices instead of constructing the feature vectors to represent the disease-miRNA pairs.

Construction of a precomputed kernel matrix has three main steps:

**Step 1**: Calculate the difference between two disease-miRNA pairs. Given two disease-miRNA pairs $d_1m_1$ and $d_2m_2$, we compute their difference $(diff\,(d_1m_1, d_2m_2))$ in three ways:

- Average approach:

$$diff\,(d_1m_1, d_2m_2) = (DisSim\,(d_1, d_2) + MiRSim\,(m_1, m_2))\,/2 \quad (3.1)$$

- Squared root approach:

$$diff\,(d_1m_1, d_2m_2) = \sqrt{(DisSim\,(d_1, d_2) \times MiRSim\,(m_1, m_2))} \quad (3.2)$$

- Center distance approach:

$$\begin{aligned} diff\,(d_1m_1, d_2m_2) =&[(DisSim\,(d_1, d_2) - AvgDisSim)^2 + \\ &(MiRSim\,(m_1, m_2) - AvgMiRSim)^2]^{1/2} \end{aligned} \quad (3.3)$$

where *DisSim* and *MiRSim* represent the similarities between diseases and miRNAs respectively. *AvgDisSim* is the average similarity of all the disease-disease pairs, and *AvgMiRSim* is the average similarity of all the miRNA-miRNA pairs. Obviously, bigger values of $diff\,(d_1m_1, d_2m_2)$ means the two pairs $d_1m_1, d_2m_2$ are more similar. Details of computing the similarities between diseases or between miRNAs are introduced in the next section.

**Step 2**: Constructing the kernel matrix for training samples. For a training set of $M$ samples $\{d_1m_1, d_2m_2, ..., d_Mm_M\}$ with class labels $\{l_1, l_2, ..., l_M\}$, the training kernel matrix, denoted as $TKM$, is given by:

$$TKM = \begin{pmatrix} k_{11} & \cdots & k_{1M} \\ \vdots & \ddots & \vdots \\ k_{M1} & \cdots & k_{MM} \end{pmatrix} \quad (3.4)$$

where, $k_{ij} = diff(d_im_i, d_jm_j)$ is the difference between the two pairs $d_im_i$ and $d_jm_j$.

**Step 3**: Constructing the kernel matrix for testing samples. For a testing set of $n$ samples $\{D_1M_1, D_2M_2, \ldots, D_nM_n\}$, the kernel matrix for the testing

samples, denoted by *PKM*, is given by:

$$PKM = \begin{pmatrix} k'_{11} & \cdots & k'_{1M} \\ \vdots & \ddots & \vdots \\ k'_{n1} & \cdots & k'_{nM} \end{pmatrix} \tag{3.5}$$

Using $TKM$ and $PKM$ as input to libSVM, the class labels of the $n$ testing samples can be predicted, and the probabilities of the predictions can be derived at the same time.

### 3.2.4 Measuring the pairwise similarities of diseases or miRNAs

Disease similarity between two diseases, denoted by *DisSim*, is measured in two parts: the disease semantic similarity (*SemSim*) and the functional similarity between disease-related gene sets (*FunSim*). The multiplication of *SemSim* and *FunSim* is defined as *DisSim*. The definition of *FunSim* is referred to the SemFunSim method (Cheng et al. 2014). We implemented the algorithm and obtained the *FunSim* measurements between 2802 diseases. The *SemSim* was computed with the R package DOSE (Yu, Wang, Yan & He 2015). For the DOSE, we applied Resnik's (Resnik et al. 1999) definition of the common ancestor for two given terms. To avoid too many zero values of the similarities, we integrated *SemSim* and *FunSim* using a sum (instead of multiplication) and a weight parameter $\alpha$. The new similarity measurement between disease $d_i$ and disease $d_j$ is computed by

$$DisSim\,(d_i, d_j) = \alpha \times FunSim\,(d_i, d_j) + (1 - \alpha) \times SemSim\,(d_i, d_j) \tag{3.6}$$

MiRNA similarity between two miRNAs, denoted by *MiRSim* is also measured in two parts: the sequence similarity (*SeqSim*) and the function similarity (*funSim*). *SeqSim* evaluates the similarity of the two miRNA sequences. We applied the idea of pseudo amino acid composition (Chou 2001) to represent a miRNA as a $(4 + \lambda)$-dimension vector. This idea was originally proposed to represent protein sequences as vectors.

Given a RNA sequence $R : r_1, r_2, \ldots r_i, \ldots r_L$, where $r_i \in \{A, G, C, U\}$. Then, $R$ is represented as a vector $V_R = [v_i]_{1\times(4+\lambda)}$, where the first four components stand for the occurrence frequencies of the 4 native nucleotides, and the latter $\lambda$ components represent the sequence order effects of the nucleotides of $R$. The $t$-th ($t < L$) tier sequence order effect $\theta_t$ is calculated by

$$\theta_t = \frac{1}{L-t}\sum_{i=1}^{L-t}\Theta\left(r_i, r_{i+t}\right) \tag{3.7}$$

$$\Theta\left(r_i, r_{i+t}\right) = \left(M_i - M_{i+t}\right)^2 \tag{3.8}$$

$$M_i = \frac{M_i^0 - \sum_{j=1}^{4}\frac{M_j^0}{4}}{\sqrt{\frac{\sum_{j=1}^{4}\left(M_i^0 - \sum_{j=1}^{4}\frac{M_j^0}{4}\right)^2}{4}}} \tag{3.9}$$

where, $M_i$ is the normalized $i$th ($i$=1, 2, 3, 4) molecular weight of the nucleotide. The original molecular weights ($M_{\mathrm{i}}^0$) of the four nucleotides are 135.1270 for A, 151.1261 for G, 111.1020 for C and 112.0868 for U. Then $V_R = [v_1, v_2, \ldots v_u, \ldots, v_{4+\lambda}]$,

$$v_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{4} f_i + w\sum_{j=1}^{\lambda}\theta_j}, (1 \le u \le 4) \\ \frac{w\theta_{u-4}}{\sum_{i=1}^{4} f_i + w\sum_{j=1}^{\lambda}\theta_j}, (5 \le u \le 4+\lambda) \end{cases} \tag{3.10}$$

In this work, we set $\lambda = 5$ and the weight factor $w = 0.05$. $f_u$ is the occurrence frequencies of the nucleotide $u$. Then, each of the miRNA sequence $R$ is represented as a 9-dimension vector $V_R = [v_i]_{1\times 9}$. Overall, the sequence similarity is given by

$$SeqSim\left(R_i, R_j\right) = 1 - \frac{SeqDis\left(V_i, V_j\right) - min(SeqDis)}{max(SeqDis) - min(SeqDis)} \tag{3.11}$$

$$SeqDis\left(V_i, V_j\right) = |V_i - V_j| \tag{3.12}$$

where, $|\cdot|$ is the Euclidean distance, and *min(SeqDis)* and *max(SeqDis)* represent the maximum value and the minimum value of all the *SeqSim* values between different miRNAs.

The *funSim* measurement is computed similarly as computing *FunSim*, namely a *funSim* between two miRNAs can be represented as the similarity between the two miRNA target sets. Similar to the measurement of *DisSim*, *MiRSim* of two miRNAs $R_i$ and $R_j$ is measured by integrating *funSim* and *SeqSim* with the same parameter $\alpha$ as follows:

$$MiRSim\,(R_i, R_j) = \alpha \times funSim\,(R_i, R_j) + (1-\alpha) \times SeqSim\,(R_i, R_j) \quad (3.13)$$

Among all the datasets we mentioned previously, 551 different mature miRNAs were involved. Thus, we obtained the similarities between these 551 miRNAs in this work (details of the miRNAs and their targets listed in Supplementary file 2). Together with the similarities between 2802 diseases (details of the disease-gene associations listed in Supplementary file 1), these plenty of similarity information provides us adequate data to investigate associations between diseases and miRNAs.

## 3.3 Results

### 3.3.1 The optimal precomputed kernel matrix and the prediction performance

There are a weight parameter $\alpha$ and a kernel matrix type $KMT$ which can be properly set to build an optimal prediction model in this work. Parameter $\alpha$ is used to mediate the similarities between diseases and the similarities between miRNAs, while $KMT$ selects a kernel matrix type for support vector machine (SVM) to make an accurate classification. Detailed explanation of $\alpha$ and $KMT$ can be found in **Methods**. Experiments for the proper selection of $\alpha$ and $KMT$ were conducted under three steps: (1) construction of training data. We extracted 1487 known disease-miRNA associations between 107 diseases and 276 miRNAs from the miR2Disease database and used them as the set of positive training samples (denoted as positive_miR). We also constructed a set of 4638 negative samples between 53 diseases and 538 miRNAs after a comprehensive analysis of the GSE

accessions (denoted as negative_expression). We randomly selected 1487 negative samples from negative_expression to construct a balanced training data set; (2) prediction model construction. This step has two layers of loops. The outer loop changes the value of $\alpha$ from 0 to 1 with a step of 0.1, while the inner loop sets $KMT = 1, 2$, or 3, which represent the three different types of kernel matrices (i.e., the average type, the squared root type and the center distance type). A prediction model was constructed with each $\alpha$ and $KMT$; (3) performance evaluation. We implemented 10-fold cross-validation on the balanced dataset with different $\alpha$ and $KMT$ and the seven performance metrics were computed. We ran the experiment 100 times. The averages of the seven indices were taken over the 100 times. **Figure 3.1** shows the AUC values and F1 scores.

The squared root type of $KMT$ outperforms the other two types. When $\alpha$ increases, the AUC and F1 score increase first but then drop down, suggesting that the integration of different types of similarities can improve the prediction performance. Furthermore, when $\alpha = 0$ or $\alpha = 1$, the average type and the squared root type can still achieve the AUC values around 0.92 and F1 scores about 0.9. It means that our precomputed kernel matrix method can have a good prediction performance even with just one kind of similarity information. Comparing the curves in **Figure 3.1**, it can be seen when $\alpha$ is around 0.8, the curves achieve better AUCs and F1 scores. Thus, we chose the squared root type of $KMT$ and set $\alpha = 0.8$ for our prediction model.

To evaluate whether our prediction performance was obtained by chance, we conducted a permutation test as Jiang et al. (Jiang et al. 2013) did. We did not use the true labels of the samples (positive samples and negative samples) but distributed the labels randomly. Then, we implemented the 10-fold cross-validation and observed the changes of the performance. The positive_miR data set was adopted as the positive samples and balanced training data sets were built. The normal predictions (true labels) were considered as the control group while the permutation tests were regarded

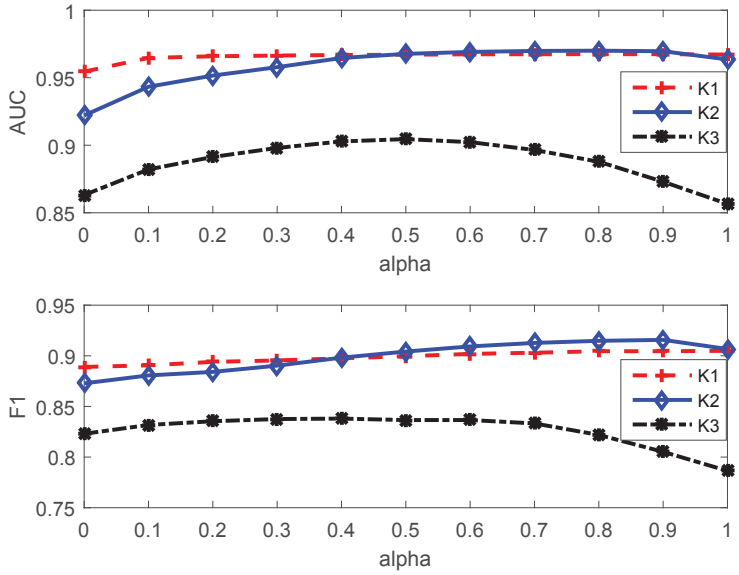


Figure 3.1: **Performances of the predictions under different precomputed kernel matrix and $\alpha$.** We mainly compare the AUC values and the F1 scores of each models with different parameters. K1, K2 and K3 represent the three kernel matrix types such as the average type, the squared root type and 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.

as the test group. All these two groups of experiments were repeated 10 times. The ROC curves of the test group and control group are shown in **Figure 3.2**. The ROC curve of the test group is nearly overlapped with the random lines while the ROC curve of the control group can achieve an AUC value of 0.97, which indicates that the performance of our prediction model was not produced occasionally but contains biological significance.

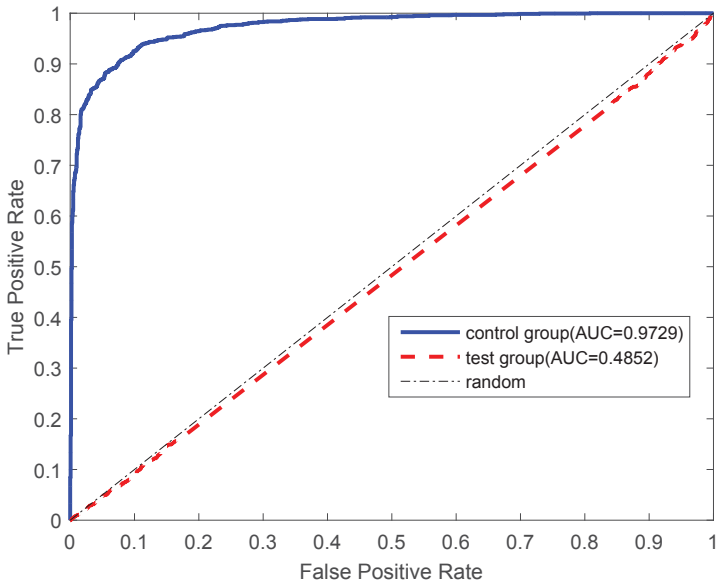


Figure 3.2: **The ROC curves of the permutation test.** The experiment includes the test group and the control group parts. The test group part used the permutated labels for the training samples while the control group part uses the original labels of the same training dataset. Both of the two parts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

### 3.3.2 Effect of the size of the negative samples on the prediction performance

To investigate whether the number of negative samples affects the performance of our predictions, we fixed the size of positive samples as the size of the positive_miR data set, and changed the number of negative samples in the training data set. All the negative samples were randomly selected from the negative_expression data set. We varied the number of negative samples from 3 times the number of positive samples to 2 times, to equal size, and to 80% of the size of positive samples, 60%, 40% and 20%. In addition, the positive samples from the positive_HMDD (totally 4041 positive samples which were extracted from the HMDD database) excluding those samples already in the

data set of positive_miR were adopted to build the validation data set. There are 3484 positive samples in this validation data set. Again, 10-fold cross-validation was implemented on the training data. The prediction model was then tested on the validation data set. As the samples in the validation data set are all positive samples, we just computed the accuracy but not other metrics. All the experiments were repeated 100 times. The average performances are depicted in **Figure 3.2** to show the changes of AUC and mcc values of the cross validation experiments and the accuracy based on the validation dataset when the size of negative samples changes (the size ratio between the negative and positive samples is displayed on the x-axis).

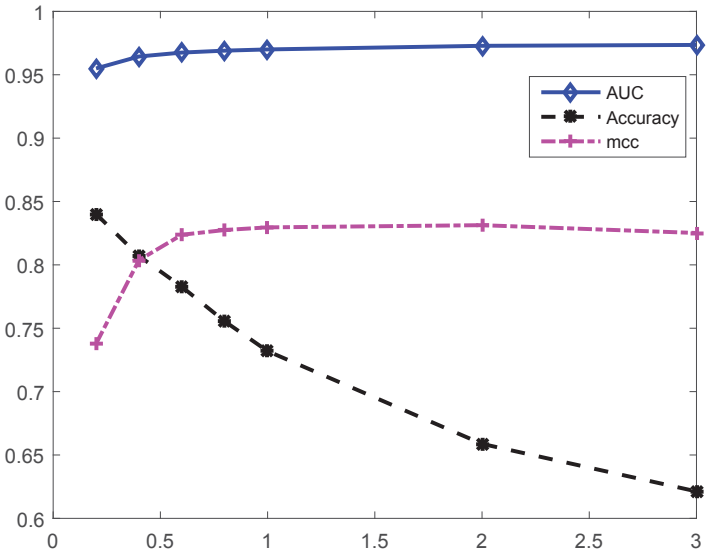


Figure 3.3: **Performances of the prediction models with different size ratio of negative and positive samples.** The prediction model was trained on the sample sets with different ratio of negative and positive samples. The x-axis shows the ratios. AUC and mcc values were computed based on 10-fold cross 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.

We can find that the AUC values have nearly no changes under different size ratios between negative and positive samples. However, the accuracy of the prediction on the validation data set drops when the size of negative samples increases. But, the mcc value increases till the size of negative samples is equal to that of positive samples. Then, it keeps at the same level even more negative samples are added. As mcc is a more comprehensive performance index than accuracy, we suggest that a balanced training data set of positive and negative samples should be adopted to infer new disease-miRNA associations as we did in this work.

### 3.3.3 Performance comparison when changing the approach of selecting negative samples

The negative samples of disease-miRNA relationship randomly selected from the negative_expression data set were used by this work for the training of the prediction model. There are other ways for the construction of negative data sets, such as by random selection from the un-connected disease-miRNA pairs. We compared the performances of our prediction model when the approach to select negative samples was changed. The positive samples were always the same, i.e., the data set positive_miR containing 1487 known disease-miRNA associations.

The negative data set formed by a random selection from those un-connected disease-miRNA pairs is named negative_random (there are total 26704 disease-miRNA pairs). We conducted two experiments. In the first experiment, we used all the 1487 positive samples from positive_miR and 1487 negative samples randomly selected from the negative_expression data set to build the training data set. The second experiment is similar to the first one with the only difference that the 1487 negative samples were randomly selected from negative_random. 10-fold cross-validation was conducted on the training data sets. To get a test performance, we also used the above validation data set to test the prediction models. All these experiments were repeated 100 times, and the average performance was taken to reduce the

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

| | 10-fold cross-validation | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| negative | specificity | recall | precision | accuracy | F1 | mcc | AUC | Accuracy |
| expression | 0.9194 | 0.9107 | 0.9191 | 0.9151 | 0.9147 | 0.8306 | 0.9704 | 0.7315 |
| random | 0.7719 | 0.7808 | 0.7746 | 0.7764 | 0.7773 | 0.5534 | 0.7315 | 0.5077 |

bias of the predictions (**Table 3.1**).

It is clear that the 10-fold cross-validation performance of selecting negative samples from negative_expression significantly outperformed another approach. For the 3484 samples of the validation data set, 73.15% of them can be correctly predicted by the model when the negative samples were selected from negative_expression, while the negative_random based model could only accurately predict 50.77% of the 3484 disease-miRNA associations. This comparison indicates that the approach for the selection of negative samples has significant impact on the prediction performance. The best choice is to select negative samples based on the analysis of expression data.

### 3.3.4 Performance comparison: prediction of disease-miRNA relationships by different methods

A number of methods have been proposed to make predictions of unknown disease-miRNA relationships. We compared the performance of our prediction method with three state-of-the-art methods: RLSMDA (Chen & Yan 2014), the method proposed by Xu et al. (Xu, Li, Lv, Li, Xiao, Shao, Huo, Li, Zou, Han et al. 2011), and Jiang's method (Jiang et al. 2013). RLSMDA is a semi-supervised method that does not need any negative samples. Xu's method is a supervised approach and it collects the negative samples according to tissue-specific and expression properties of the miRNAs. Jiang's method is also a supervised method. It has utilized a set of 270 negative samples randomly selected from the un-connected disease-miRNA pairs of a miRNA-disease bipartite network. More detail of these procedures and the results can be found in the contents and Supplementary file 5: Figure S1-S3. The source

Table 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.

| Methods | sample size | cv type | Specificity | Sensitivity | Accuracy | AUC |
|---|---|---|---|---|---|---|
| RLSMDA | 1184+ | LOOCV | – | – | – | 0.9475 |
| our model | 1184+,1184- | LOOCV | 0.9367 | 0.9368 | 0.9367 | **0.9896** |
| Xu's method | 37+, 44- | 5-fold | 0.8833 | 0.8643 | 0.8772 | 0.9189 |
| our model | 37+, 37- | 5-fold | **0.9990** | **1.000** | **0.9995** | **0.9854** |
| Jiang's method | 270+, 270- | 10-fold | 0.9125 | 0.7338 | 0.8232 | 0.8884 |
| our model | 263+, 263- | 10-fold | **0.9274** | **0.8982** | **0.9128** | **0.9871** |

codes of these literature methods were not available. We implemented the RLSMDA algorithm, but not the complicated Xu's or Jiang's method. For a fair comparison, their data sets and performance metrics were exactly used by our method. The performance metrics are: specificity, recall (or sensitivity), precision, accuracy, and AUC (area under the ROC curve). More details of the implementation and data sets are described in Supplementary file 5, the positive samples are listed in Supplementary file 7.

The AUC performances are benchmarked in **Table 3.2**. The ROC curve of our method is depicted in **Figure 3.4** in comparison with the curve of the RLSMDA method under the same data set and the same leave-one-out cross-validation (LOOCV). The ROC curves for the comparison of our methods with all the three methods are also showed in Supplementary file 5: Figure S4-S6. Our prediction model achieves much better AUC values than the three state-of-the-art methods. The superior performance of our prediction method is mainly attributed to the careful selection of reliable negative samples as well as the precomputed kernel matrix which can identify more positive samples.

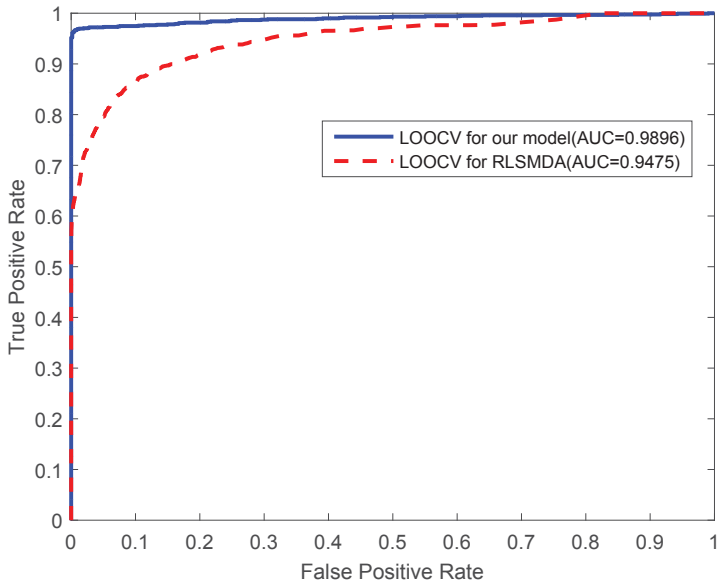


Figure 3.4: **The ROC curves of our model compared with RLSMDA based on the same positive samples.** The comparison is based on the same positive sample set and the different prediction model of RLSMDA and our newly designed model. The average AUC value of our model is 0.9896 while the RLSMDA obtains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

### 3.3.5 The predicted miRNAs that are related to breast and prostate cancer: Case studies

In this section, we report details of the predicted miRNAs which are likely related to breast cancer and prostate cancer. Breast cancer is the leading type of cancer in women, accounting for 25% of all women cancer patients (Stewart & Wild 2014). Prostate cancer is the second most common type of cancer and the fifth leading cause of cancer-related death in men (Stewart & Wild 2014). We have taken the following three steps for this case study: (1) the prediction model was trained on the RLSMDA data set of disease-miRNA associations (Chen & Yan 2014) which contains 1184 disease-miRNA associations; (2) the prediction model was applied to make predictions for those disease-miRNA pairs whose relationships were unknown in this

data set; (3) the positively predicted disease-miRNA pairs were evaluated using the latest version of databases such as miRCancer (Xie et al. 2013), miR2Disease (Jiang et al. 2009) and HMDD (Lu et al. 2008), which stores newer disease-miRNA associations than the RLSMDA data set does. In fact, the RLSMDA data set stores only 78 and 34 miRNAs associated with breast cancer and prostate cancer respectively. However, the latest version of the three databases stores 227 and 152 miRNAs which have been found related to breast and prostate cancer. Thus, our predicted results can be fairly verified by the literature ground truth. As some of the predicted disease-miRNA associations were not covered by the three databases, we also searched other web resources to confirm the prediction results.

We constructed 100 prediction models (for making reliable predictions), each time using all the 1184 disease-miRNA pairs as the positive samples and a set of randomly selected 1184 negative samples from the negative_expression data set (a data set of 4638 negative samples based on the analysis of expression data). If an unknown cancer-miRNA relationship is positively predicted by all the 100 models, then a strong association exists between the cancer and the miRNA. The association probabilities derived by the 100 models are averaged to indicate the strong association. **Figure 3.5** shows the 30 top-ranked positively predicted miRNAs related to breast and prostate cancer in terms of the average probabilities of the 100 models for the miRNAs. The edges at the (a) part represent the breast cancer-miRNA associations while the edges at the (b) part show the prostate cancer-miRNA associations. The labels on these edges represent the ranking positions and evidence type of the prediction results. The characters "*", "#" or "$" stand for that the corresponding associations can be confirmed by the records in the miR2Disease database, the HMDD database or the miRCancer database respectively. The character "@" means that the association can be confirmed by other articles. Otherwise, the predicted associations could not be confirmed to our best knowledge. Overall, 58 of the 60 predicted disease-miRNA relationships can be verified by the newer databases or by other

literature work.

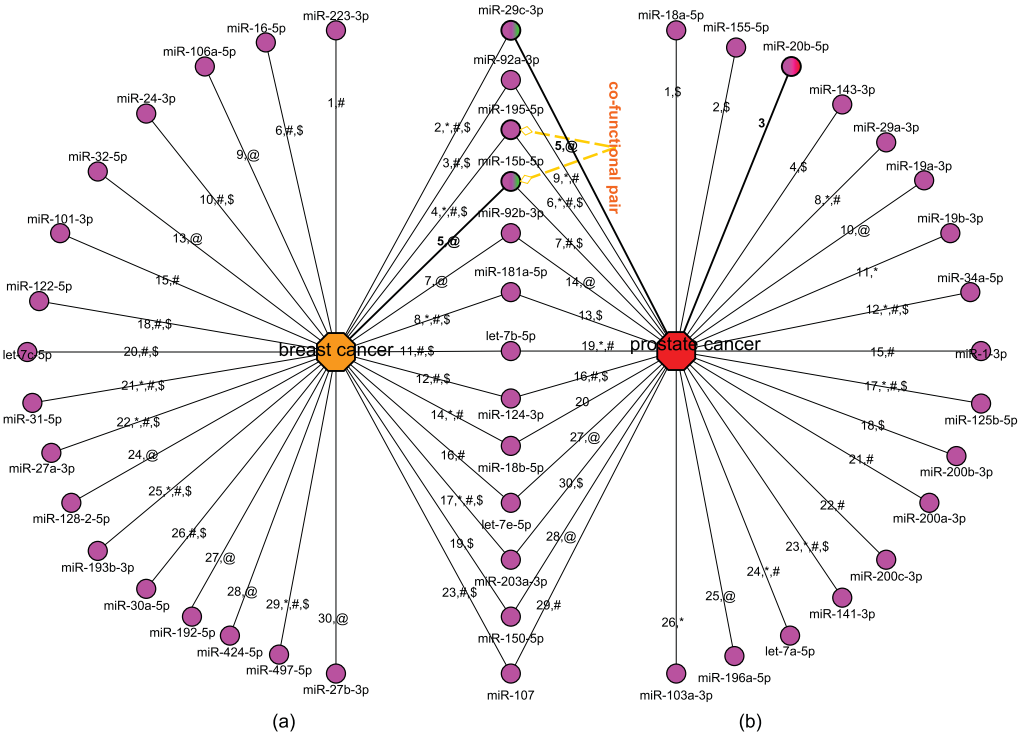


Figure 3.5: **The top 30 predicted breast cancer-miRNA and prostate cancer-miRNA associations and the verification resources.** The left part shows the predicted breast cancer related miRNAs and the right part gives the predicted prostate cancer related miRNAs. The labels of the edges illustrate the 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.

**Figure 3.6**(a) shows the percentages of the predicted disease-miRNA associations that can be verified when the number of top-ranked miRNAs varies from 10 to 150. The x-axis is the number of predictions ($\times$ 10) while the y-axis is the percentages of the verified predictions. For the first 10

to 50 predicted miRNAs associated with breast cancer or prostate cancer, 100% and 96% of them can be verified by the three newer databases or literature. The percentages drop to 98% and 88% when we assess on the first 100 predicted associations. This indicates that a more reliable predicted disease associated miRNAs can be ranked at a higher position by our method.

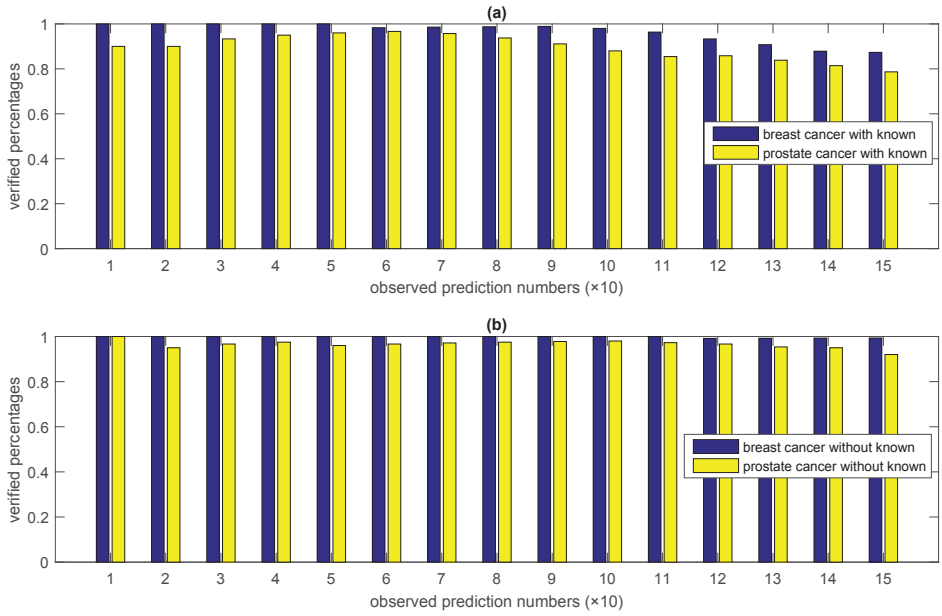


Figure 3.6: **The percentages of the predicted disease-miRNA associations that can be verified.** Panel (a) introduces the prediction performance of the model with the known cancer (breast and prostate cancer) related miRNAs. Panel (b) shows the prediction performance after the removal of the existing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

A novel association predicted by our method is about hsa-miR-15b (mapped as hsa-miR-15b-5p by miRBase) and breast cancer. Hsa-miR-15b is ranked as the 5th leading breast cancer related miRNA. This miRNA is an epidermal growth factor induced miRNA, and its association with breast

cancer has not been recorded by any existing databases. However, a new discovery in 2015 can verify that there is an inverse correlation between the high expression of miRNA-15b and the low expression of its target gene MTSS1 in the tissues of breast cancer patients with the aggressive basal subtype (Kedmi, Ben-Chetrit, Körner, Mancini, Ben-Moshe, Lauriola, Lavi, Biagioni, Carvalho, Cohen-Dvashi et al. 2015). The growth factor-inducible miRNAs can mediate the mechanisms underlying the progression of breast cancer. Another novel association predicted by our method is about hsa-miR-29c (mapped as hsa-miR-29c-3p by miRBase) and prostate cancer. This association is also ranked at the 5th position among the predicted prostate cancer related miRNAs, but it has not been recorded by any existing databases. A recent report claimed that miR-29c together with other five miRNAs such as miR-29a, miR-29b, miR-26a, miR-26b and miR-218 can control the expression of metastasis-promoting LOXL2 gene during the development of prostate cancer (Kato, Kurozumi, Goto, Matsushita, Okato, Nishikawa, Fukumoto, Koshizuka, Ichikawa & Seki 2017).

For the association miR-20b-prostate cancer which cannot be verified, we found that Moltzahn et al. (Moltzahn, Olshen, Baehner, Peek, Fong, Stöppler, Simko, Hilton, Carroll & Blelloch 2011) has reported the upregulation of miR-20b in prostate cancer patients comparing with the healthy samples according to their robust multiplex qRT-PCR method profiling. However, this upregulation was not statistically significant based on the follow-up PCR. This clue shows that there may be an association between miR-20b and prostate cancer.

For some diseases, there have no currently known associations with any miRNAs. To test whether our prediction algorithm is still applicable for such situations, we conducted another experiment. In the experiment, we removed all the known miRNA associations with breast cancer or prostate cancer from the RLSMDA data set. The objective was to see whether our model can correctly predict these purposely removed and currently known breast cancer-miRNA or prostate cancer-miRNA associations. The

prediction results are shown in **Figure 3.6**(b). Our model has a superior performance for predicting disease-miRNA associations even when there is no known association for these two cancers. Of the top 50 predicted disease-miRNA associations, all the predicted breast cancer-miRNA associations can be confirmed by the existing databases or literature, while 96% of the top 50 predicted prostate cancer-miRNA associations can be confirmed. The confirmation rates for the top-100 predicted associations can still maintain at a very high level. Moreover, the breast cancer-hsa-miR-15b-5p and the prostate cancer-hsa-miR-29c-3p can still be predicted and ranked highly. More details of the predicted and verified disease-miRNA pairs can be found in Supplementary file 5: Table S3-S6. The code in the Supplementary file 6 which implements our prediction algorithm has a default setting to output no more than 100 miRNAs for a given disease.

## 3.4 Conclusion

In this chapter, we designed a precomputed kernel matrix SVM method with reliable negative samples to address the disease-miRNA association prediction question. This method solves part of my research question **Q1** (see **Section 1.2**). The contributions are concluded as follows (also described in **Section 1.3 C1**): (1) We proposed a new idea for selecting reliable negative samples of disease-miRNA relationship which can overcome the problem of lacking negative samples for machine learning methods to make reliable predictions of disease-associated miRNAs; (2) We applied the miRNA sequence information to compute miRNA similarities. Various ways for computing disease similarities and miRNA similarities were integrated to improve the prediction performances; (3) Our prediction model does not need to do feature selection, and it is applicable for large-scale prediction of disease-associated miRNAs; (4) Our prediction model can work well for those miRNAs that have no currently known miRNA-disease associations.

# Chapter 4

# Cross Disease Analysis of Co-functional microRNA Pairs on A Reconstructed Network of Disease-gene-microRNA Tripartite

## 4.1 Introduction

As mentioned in previous Section 1.1.2, abundant validated disease-miRNA associations have been collected by the databases such as HMDD (Lu et al. 2008) and miR2Disease (Jiang et al. 2009). In Chapter 3, a precomputed kernel matrix SVM method was introduced for new disease-miRNA association prediction. By observing those known associations and the newly predicted ones, we can find many overlapped miRNAs that associate with different diseases. This inspired us to investigate the co-functional roles of miRNAs in multi-diseases.

Pairs of miRNAs have been reported to work cooperatively for regulating an individual gene or a cohort of genes that participate in similar processes (Lai,

Schmitz, Gupta, Bhattacharya, Kunz, Wolkenhauer & Vera 2012, Xu, Li, Li, Li, Shao, Bai, Chen & Li 2013). This cooperativity (or co-function) is a frequent regulation mechanism of miRNAs for an enhanced target repression which has exhibited distinctive and fine-tuned target gene expression patterns (Schmitz, Lai, Winter, Wolkenhauer, Vera & Gupta 2014). Investigation on miRNA cooperativity can help systematically understand miRNA functions (Xu, Li, Li, Lv, Ma, Shao, Xu, Wang, Du, Zhang et al. 2011) and study their potential disease links (Xiao, Xu, Guan, Ping, Fan, Li, Zhao & Li 2012).

Using miRNAs as diagnostic and therapeutic targets, miRNA therapeutics is a promising research area that designs sophisticated strategies to restore or inhibit miRNA expression for the treatment of cancer and other diseases. For example, a therapy with the vector-encoded pair miR-15a and miR-16-1 has been proposed for the treatment of patients with chronic lymphocytic leukaemia (CLL) (Ling et al. 2013). The microRNA cluster miR-216a/217 was also reported to target genes PTEN and SMAD7 to induce the epithelial-mesenchymal transition, which finally promotes the drug resistance and recurrence of liver cancer (Xia, Ooi & Hui 2013). Such co-functional miRNA pairs are more suitable as drug targets instead of using individual ones. Large scale detection of novel co-functional miRNA pairs is an important pre-step to identify proper miRNA pairs as more effective drug targets. Currently, abundant disease-gene association information is stored in Online Mendelian Inheritance In Man (OMIM) (Hamosh, Scott, Amberger, Bocchini & McKusick 2005) and Comparative Toxicogenomics Database(CTD) (Davis, Murphy, Johnson, Lay, Lennon-Hopkins, Saraceni-Richards, Sciaky, King, Rosenstein, Wiegers et al. 2013); disease-miRNA associations are recorded in miR2Disease and HMDD; and miRNA-target regulations are recorded in miRecords (Xiao et al. 2009) and miRTarBase (Hsu et al. 2010). Linking and integrating these databases, it can be inferred which diseases are correlated with the same genes or with the same miRNAs, and which miRNAs have the same target disease genes. Our hypothesis is that some of the miRNAs

can regulate their common targets cooperatively and have roles in the development of a series of diseases.

The focus of this work is on the detection and prioritization of multi-disease associated co-functional miRNA pairs. A multi-disease associated co-functional miRNA pair is a pair of miRNAs whose common target genes are associated with a series of diseases. Here, the definition of co-function for the miRNA pairs is broader than the definition of cooperativity as proposed in (Broderick, Salomon, Ryder, Aronin & Zamore 2011, Moore, Scheel, Luna, Park, Fak, Nishiuchi, Rice & Darnell 2015). **Figure 4.1** shows an example of multi-disease associated co-functional miRNA pairs detected from a disease-gene-miRNA (DGR) tripartite network. From this example, we can see that multi-disease associated co-functional miRNA pairs may hold a vast mechanism underlying multiple disease development, similarly like the basic cellular functions maintained by housekeeping genes. More importantly, these miRNAs can be considered as the common drug targets of these diseases for the design and development of multi-purpose drugs.

MiRNA co-function mechanisms have attracted intensive research recently (Yoon & De Micheli 2005, Xu, Li, Li, Lv, Ma, Shao, Xu, Wang, Du, Zhang et al. 2011, Xiao et al. 2012, Wu, Li, Zhang, Yao, Wu, Han, Liao, Xu, Lin, Xiao et al. 2013), with the focus on the analysis of miRNA-target networks or on the analysis of disease-miRNA associations for a specific disease. Our work advances the current research with two steps: (i) We reconstruct a DGR tripartite network through the integration of existing databases with our newly predicted disease-miRNA associations, and (ii) we propose a novel scoring method to prioritize the potential multi-disease associated co-functional miRNA pairs.

Combining our predicted disease-miRNA associations (by the proposed precomputed kernel matrix SVM method described in Chapter 3) with those literature-maintained associations between diseases, miRNAs and genes, we construct a more complete DGR tripartite network to detect and prioritize multi-disease associated co-functional miRNA pairs. Given a miRNA pair,

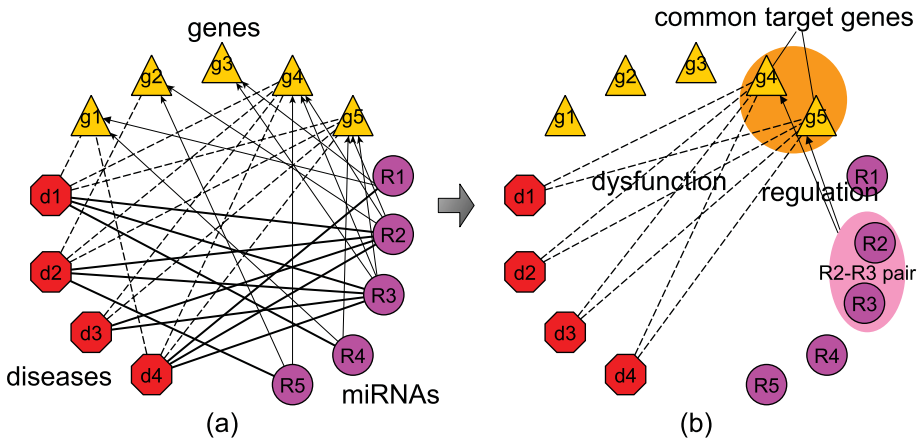


Figure 4.1: **An example: From a DGR tripartite network to a co-functional miRNA pair.** The network in panel (a) contains known associations among the genes g1, g2, g3, g4, and g5, the diseases d1, d2, d3, and d4, and the miRNAs R1, R2, R3, and R4. In this example, miRNAs R2 and R3 are both associated 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.

our scoring method *cfscore* considers the function relationship between the two miRNAs, the co-dysexpression of the two miRNAs in the disease tissues and the relationship between the common target genes and the associated diseases of these miRNAs. We are also interested in finding the exact targets dysregulated by the co-functional miRNA pair during the diseases' development. We call them the co-functional targets of the co-functional miRNA pair. The flowchart of our work is described in **Figure 4.2**.

This method was tested on the cancer and non-cancer disease related DGR tripartite networks. The top 50 multi-disease associated co-functional miRNA pairs were concentrated for deep analysis. We found that most

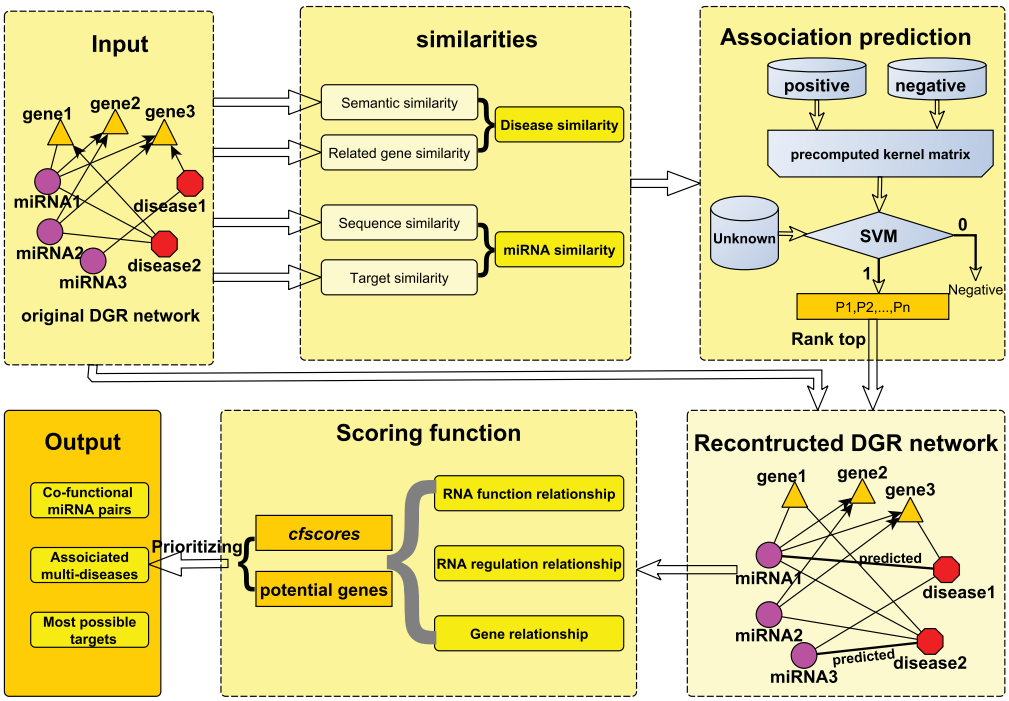


Figure 4.2: **The flowchart of our prediction and scoring method.** Our work includes the parts such as material collection, similarity computing, association prediction, network reconstruction, scoring and priorization of the co-function miRNA pairs and result output.

of them were from the same miRNA families or miRNA clusters. The comparison of the co-functional pairs from the two DGR networks suggests that the dysregulation mechanisms of miRNAs in the cancers are more complex. It has also been shown that the analysis of multi-disease associated co-functional miRNAs can help understand the regulation mechanisms of miRNAs in the development of different diseases and thus can provide new knowledge for the diagnosis or treatment of the diseases.

## 4.2 Method

Our method for the detection and prioritization of co-functional miRNA pairs and cross disease analysis includes three main computational steps: (i) Reconstructing the DGR tripartite network by combining the known relationships of diseases, miRNAs and genes with those predicted disease-miRNA associations, (ii) Ranking the candidate co-functional miRNA pairs via a novel scoring method, (iii) Determining the potential co-functional target genes of these co-functional miRNA pairs. Details of these steps are described in the following subsections. Those data sets that we adopted have been described in the **Method** section of Chapter 3.

### 4.2.1 Reconstructing the DGR tripartite network

A DGR tripartite network is used to reflect the relationships between diseases, genes and miRNAs. For a disease group $V_d = \{d_1, d_2, \ldots, d_a, \ldots, d_M\}$, a gene group $V_g = \{g_1, g_2, \ldots, g_b, \ldots, g_N\}$ and a miRNA group $V_r = \{r_1, r_2, \ldots, r_c, \ldots, r_P\}$, their associations are stored in $E_{dg}$ for the disease-gene association, $E_{dr}$ for the disease-miRNA association and $E_{rg}$ for the miRNA-target association, respectively. The DGR tripartite network can be constructed by defining $d_a$, $g_b$ and $r_c$ as network nodes and connecting those nodes that there are known associations between them. For example, we connect the nodes $d_a$ and $g_b$ if their association $d_a - g_b$ is in $E_{dg}$.

In Chapter 3, we proposed a precomputed kernel matrix SVM method to predict reliable disease-miRNA associations. We have proved its excellent performance by various tests and comparisons. Those newly detected disease related miRNAs are represented by an association set $E_{dr}^{pre}$. Merging $E_{dr}^{pre}$ and previous $E_{dr}$ to produce a new set $E_{dr}^{mer} = E_{dr}^{pre} \cup E_{dr}$, we then can reconstruct the DGR tripartite network with the updated associations.

### 4.2.2 Scoring the multi-disease associated co-functional miRNA pairs

This work defines a multi-disease associated co-functional miRNA pair as a pair of miRNAs that can dysregulate the same gene or whose target genes are involved in the same cellular processes to participate in the development of a series of diseases. Such a miRNA pair has three good properties: (i) the members function cooperatively, which means they prefer to share the same targets; (ii) the members are associated with the development of a same set of diseases; and (iii) the common miRNA targets of the two miRNAs are potentially to be the common disease genes of their related diseases. These three properties can be examined on a DGR tripartite network containing various associations between miRNAs, diseases and genes.

Let $dgr = (V_d \cup V_g \cup V_r, E)$ be a DGR tripartite network, where $V_d$ is a set of diseases, $V_g$ is a set of disease genes, $V_r$ is a set of disease-related miRNAs, and $E$ is the associations between these diseases, genes, and miRNAs. Given a pair of miRNAs $R_1$ and $R_2$, $R_1, R_2 \in V_r$, we find the gene sets $G_1 = \{g_{11}, g_{12}, \ldots, g_{1k}, \ldots, g_{1m}\}$ and $G_2 = \{g_{21}, g_{22}, \ldots, g_{2t}, \ldots, g_{2n}\}$, where $g_{1k}, g_{2t} \in V_g$ and the edges $(R_1, g_{1k})$, $(R_2, g_{2t}) \in E$. We also find two subsets of diseases $D_1 = \{d_{11}, d_{12}, \ldots, d_{1p}, \ldots, d_{1x}\}$ and $D_2 = \{d_{21}, d_{22}, \ldots, d_{2q}, \ldots, d_{1y}\}$, such that $d_{1p}, d_{2q} \in V_d$ and the edges $(R_1, d_{1p})$, $(R_2, d_{2q}) \in E$. Then, for each disease $d_l$ in $D_1$ and $D_2$, we can get its related genes $d_l^g = \{g_1, g_2, \ldots, g_l, \ldots, g_z\}$.

We quantify (i) the function relationship between a pair of miRNAs, (ii) miRNA regulation relationship in different diseases, and (iii) the relationship between the shared targets of two miRNAs and the common disease genes of these miRNAs associated diseases:

- MiRNA function relationship. A function relationship between $R_1$ and $R_2$ is quantified as the proportion of the shared targets ($psg(R_1, R_2)$), namely,

$$psg(R_1, R_2) = \frac{G_1 \cap G_2}{G_1 \cup G_2} \tag{4.1}$$

- MiRNA regulation relationship in different diseases. The idea is that those miRNAs that have significant differential expression levels in different disease are more likely to function cooperatively. The co-dysexpression rate of $R_1$ and $R_2$, $rd(R_1, R_2)$, is defined with consideration of their shared diseases and the percentage of the shared diseases comparing with all the diseases in *dgr* (i.e., $|V_d|$):

$$rd(R_1, R_2) = \frac{D_1 \cap D_2}{D_1 \cup D_2} \cdot \frac{D_1 \cap D_2}{|V_d|} \tag{4.2}$$

- The relationship between the shared targets of $R_1$ and $R_2$ and the common disease genes of $R_1$ and $R_2$ shared diseases is defined as $psgc(R_1, R_2)$. The idea is that those co-functional miRNAs always co-dysregulate the common disease genes to contribute to the disease development.

$$psgc(R_1, R_2) = \frac{\bigcup_{l=1}^{s}((G_1 \cap G_2) \cap d_l^g)}{G_1 \cap G_2} \tag{4.3}$$

where $s$ is the number of diseases that the R1 and R2 shared.

The score for weighting the probability of the pair $R_1$ and $R_2$ to be a multi-disease associated co-functional pair (*cfScore*) is defined as:

$$cfScore(R_1, R_2) = psg(R_1, R_2) \cdot rd(R_1, R_2) \cdot psgc(R_1, R_2) \tag{4.4}$$

MiRNA pairs related to bigger number of diseases are more likely to reflect the general regulation mechanism. Thus, a threshold is set to control the number of diseases that the pair is associated with. There is no reliable data set for us to select an optimal threshold, we just set the threshold to be 10. We can then rank all the candidate co-functional miRNA pairs according to their *cfScores*. A higher position indicates the pair is more likely to be a multi-disease associated co-functional miRNA pair.

### 4.2.3 Determining the potential co-functional target genes

Usually, the two members of a co-functional miRNA pair can share more than one common targets. However, only part of them are really dysregulated by the miRNA pair during the development of the diseases (called the co-functional targets of this co-functional miRNA pair). As all those miRNAs shared targets can be candidate co-functional target, a probability is estimated for the candidate co-functional targets to be the exact dysregulated genes during the diseases' developments. The idea is that the candidate co-functional targets being the disease genes for more of the miRNA pair associated diseases are more likely to be the real ones. We calculate the probability of gene $g_i$, $p(g_i)$, to be a co-functional target by:

$$p(g_i) = \frac{C_{g_i \cap (D_1 \cap D_2)}}{C_{(D_1 \cap D2)}} \tag{4.5}$$

where $C_{D_1 \cap D_2}$ is the number of common diseases associated with miRNA $R_1$ and $R_2$, while $C_{g_i \cap (D_1 \cap D_2)}$ is the number of diseases associated with gene $g_i$.

## 4.3 Results

### 4.3.1 Multi-disease associated co-functional miRNA pairs and their common dysfunctional target genes

Two cancer-gene-miRNA tripartite networks were constructed to investigate the performance of our method for detecting and ranking multi-cancer associated co-functional miRNA pairs. We have merged the miRCancer database (Xie et al. 2013) with miR2Disease (Jiang et al. 2009) and HMDD (Lu et al. 2008), and collected 3655 cancer-miRNA associations between 83 cancers and 503 miRNAs. Connecting these miRNAs and

diseases to their associated genes, the first cancer-gene-miRNA tripartite network was constructed. Then, all the 3655 cancer-miRNA associations (as positive samples) and a balanced set of 3655 negative samples of cancer-miRNA association in this tripartite network were used together to train our prediction model (the precomputed kernel matrix SVM method, see Chapter 3) for inferring new cancer-miRNA associations. The prediction model was applied to all the un-connected disease-miRNA pairs between the 83 cancers and 503 miRNAs to predict whether some of them have associations or not. When a pair was predicted to have an association between a cancer and a miRNA, a probability was also estimated. A total of 3000 top-ranked associations were added to the first cancer-gene-miRNA tripartite network to form the second cancer-gene-miRNA tripartite network (i.e., a reconstructed network by adding the predicted cancer-miRNA associations). Those associations can be found in the Supplementary file 8.

On average, the 503 miRNAs are associated with 7 or 13 cancers respectively for the first and the reconstructed network; and there are 2532 and 5634 miRNA pairs in these two networks that have a *cfScore* larger than 0 and that are associated with at least 10 cancers. There are very few literatures which prove the miRNA pairs co-function during the development of more than 10 different diseases. To understand whether these miRNA pairs co-function during the development of some of the diseases, we manually searched and examined relevant literature to confirm that the individual miRNAs in the pairs can function cooperatively to regulate the same targets. Of the top-ranked 50 miRNA pairs from our reconstructed network, 40 pairs can be validated to be co-functional pairs by the literature, in comparison with 35 of the top 50 pairs from the the first tripartite network. Here, we can just confirm these pairs of miRNAs are co-functional miRNA pairs but not multi-disease associated co-functional ones. We can't find literature that discuss the relationship between miRNAs and a series of diseases simultaneously. This also implies that the addition of the predicted disease-miRNA associations into the tripartite network is useful and effective for the

study of co-functional miRNA pairs.

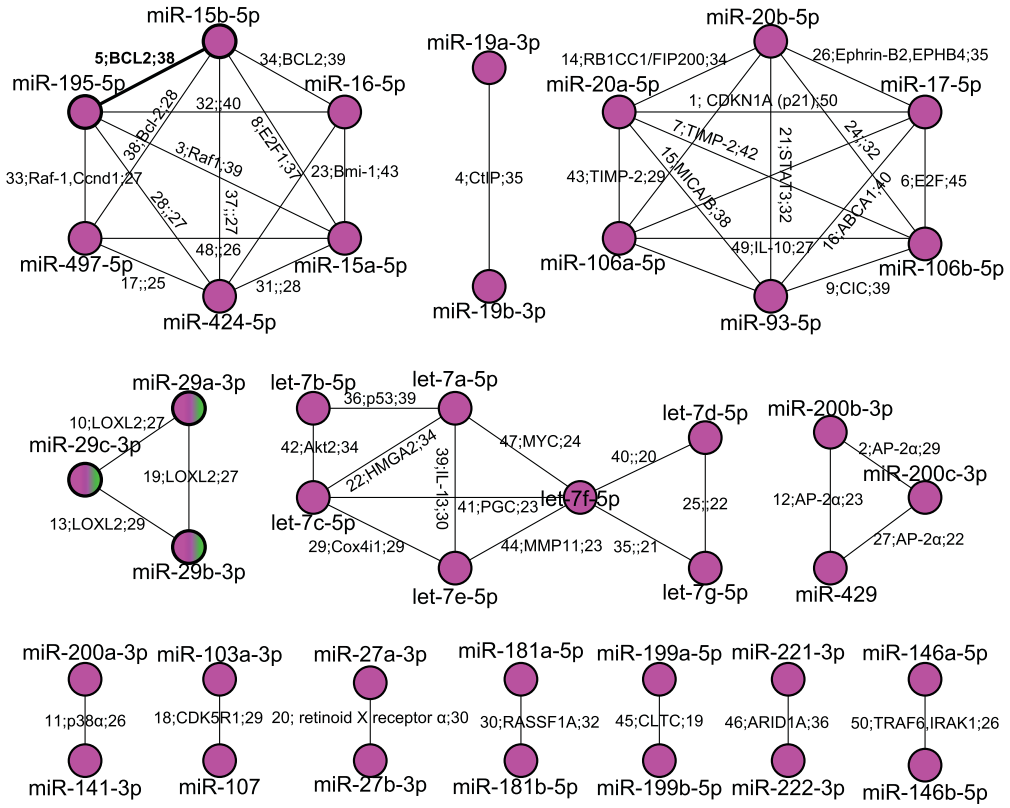


Figure 4.3: **The 50 top-ranked co-functional miRNA pairs from the reconstructed cancer-miRNA-gene network.** The labels along the edges illustrate the co-function information of the miRNAs. The first number of each label is the rank of the corresponding pair according to our prioritization method. The 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.

Details of the 50 miRNA pairs are shown in **Figure 4.3**, where on the label of each edge, the first number represents the ranking position of the

miRNA pair. If the rank number is followed by one or more gene names, it represents that the miRNA pair is a co-functional pair and has validated common targets. The number at the end of the label is the number of diseases that may associate with this co-functional pair. These multi-cancer associated co-functional miRNA pairs are mostly from the same clusters or families such as from the let-7 family (let-7a∼7e and miR-98) and the miR-17∼92 cluster (miR-17-3p, miR-17-5p, miR-18a, miR-19a, miR-19b, miR-20a and miR-92). It has been known that clustered miRNAs or those miRNAs from the same family are evolved from a common ancestor and can target functionally related genes (Hausser & Zavolan 2014). Thus, it can be easily understood that miRNAs from the same cluster or family have similar functions and can always function cooperatively. It should be mentioned that not all those miRNAs in the same families or clusters will co-function with each other. The target genes of those miRNAs from the same families or clusters are not completely overlapped. Moreover, some miRNAs that belong to different families or clusters may also be co-functional miRNAs. The 17th-ranked pair miR-497-5p-miR-424-5p is also prioritized to be a co-functional miRNA pair. However, according to the miRBase, miR-424-5p is a member of mir-322 gene family while miR-497-5p stems from the mir-497 family. They are also not clustered miRNAs.

To prove that each of those top-ranked miRNA pairs contains two co-functional miRNAs and it associates with multi-diseases, we take the 5th-ranked pair, miR-15b and miR-195, as an example. They both belong to the miR-15 family, and both of them can target gene BCL2, an important apoptosis inhibitor. In fact, this pair of miRNAs can work together with another miRNA (miR-16) in the regulation (Liu, Yang, Xie, Ren, Sun, Zeng & Sun 2012). We hypothesize that this co-functional pair may dysregulate their targets cooperatively, leading to the development of 38 different cancers such as prostate cancer (DOID:10283), prostate carcinoma (DOID:10286), stomach cancer (DOID:10534), and breast cancer (DOID:1612). The top three potential common targets of this miRNA pair are genes BCL2 (entrez

id:596), CDKN1A (entrez id:1026), and CCND1(entrez id:595). We have verified that these three genes are individually related to most of (68%, 68% or 66%) the 38 cancers. Furthermore, these three genes are all involved in four KEGG (Kanehisa & Goto 2000) pathways: hsa05215: Prostate cancer (p-value=1.5E-4), hsa05206: MicroRNAs in cancer (p-value=1.7E-3), hsa04151: PI3K-Akt signaling pathway (p-value=2.5E-3) and hsa05200: Pathways in cancer (3.2E-3) as revealed by the DAVID functional annotation tool (Huang, Sherman & Lempicki 2009*b*, Huang, Sherman & Lempicki 2009*a*). Moreover, the three genes all have the functions of the cellular response to DNA damage stimulus (GO:0006974, p-value=1.4E-4) and response to drug (GO:0042493, p-value=4.0E-4), which are important functions for the normal cells. Based on these analysis and evidences, it is suggested that the pair of miR-15b and miR-195 may contribute to the development of all the 38 different types of cancers via a similar regulation mechanism. More details of the discovered miRNA pairs and references are listed in Supplementary file 5.

We were also interested in the problem of whether the co-functional phenomenon for the non-cancer disease related miRNAs is the same as those of cancers. Thus, we constructed a non-cancer disease related DGR tripartite network containing 1625 non-cancer disease-miRNA associations between 334 miRNAs and 174 diseases extracted from the three existing databases and also containing 1625 predicted associations (Supplementary file 8). There were just 13 multi-non-cancer-disease associated co-functional miRNA pairs having a *cfscore* bigger than 0 and associating with no less than 10 different diseases. Again, we manually examined these candidate co-functional miRNA pairs. We found that 11 of them can be validated with strong evidence from literature (Supplementary file 2). Furthermore, 5 of the 13 pairs overlap with the cancer related top 50 miRNA pairs. This indicates that the co-functional mechanism exists not only for the cancer related miRNAs but also for non-cancer disease related miRNAs. (However, there are less multi-non-cancer disease associated co-functional pairs comparing

with cancers.) The dysregulation mode of the specific miRNAs for the development of cancers and the non-cancer diseases may shed a light to some extent.

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

| cancer related co-functional miRNA pairs | | | | |
|---|---|---|---|---|
| miRNA1 | miRNA2 | rank | cancer numbers | co-functional targets |
| miR-15a-5p | miR-15b-5p | 8 | 37 | BCL2; CDKN1A; CCND1; VEGFA; MTHFR; IFNG; FGF2; FGFR4; SMAD7; CHEK1 |
| miR-17-5p | miR-20a-5p | 1 | 50 | TP53; CCND1; BCL2; CDKN1A; MDM2; VEGFA; MYC; HIF1A; CXCL8; SOD2 |
| miR-29a-3p | miR-29b-3p | 19 | 27 | BCL2; MDM2; VEGFA; CASP8; MMP2; PTEN; AKT2; SPARC; VHL; DNMT3B |
| miR-29a-3p | miR-29c-3p | 10 | 27 | BCL2; MDM2; VEGFA; CASP8; MMP2; PTEN; AKT2; SPARC; VHL; DNMT3B |
| miR-29b-3p | miR-29c-3p | 13 | 29 | BCL2; MDM2; VEGFA; CASP8; MMP2; PTEN; VHL; AKT2; SPARC; CCNA2 |
| non-cancer diseases related co-functional miRNA pairs | | | | |
| miRNA1 | miRNA2 | rank | disease numbers | co-functional targets |
| miR-15a-5p | miR-15b-5p | 5 | 10 | IFNG; MTHFR; RARB; BCL2; CSNK1E; JARID2; PDCD1; ALDH3B1; APP; CDC25A |
| miR-17-5p | miR-20a-5p | 2 | 17 | CXCL8; SOD2; BCL2; ESR2; TP53; VEGFA; F3; ITGA2; PTGER4; CCL5 |
| miR-29a-3p | miR-29b-3p | 1 | 20 | MMP2; VEGFA; COL3A1; BCL2; FGB; CASP8; FGA; S100B; SPARC; TGFB3 |
| miR-29a-3p | miR-29c-3p | 4 | 13 | MMP2; COL3A1; VEGFA; AKT2; CASP8; FGB; MDM2; SGK1; TET2; BCL2 |
| miR-29b-3p | miR-29c-3p | 3 | 14 | MMP2; COL3A1; VEGFA; AKT2; CASP8; FGB; MDM2; MMP15; SGK1; MMP24 |

### 4.3.2 An in-depth analysis of five co-functional miRNA pairs

To further understand the regulation mechanism of the co-functional miRNA pairs, we particularly focused on the potential common targets of the 5

overlapping co-functional pairs (**Table 4.1**). The first two columns list the two individual miRNAs in the co-functional miRNA pairs, the third column shows the number of diseases that may relate to the miRNA pairs, and the last column lists the co-functional targets of these miRNA pairs which are related to multiple diseases. Here, a target gene is ranked higher if it relates to more diseases. It can be seen that even though there are common co-functional miRNA pairs between cancers and non-cancer diseases, the co-functional targets of these miRNA pairs are different from each other. For example, for the two miRNA pairs that both are members of the miR-15 family (miR-15a/b), the top three possible co-functional targets for the non-cancer diseases are IFNG, MTHFR, RARB, while for cancers are BCL2, CDKN1A and CCND1. Meanwhile, there are a lot of genes repeatedly relate to various miRNA pairs such as the last three miRNA pairs from **Table 4.4**. Thus these miRNA pairs may function cooperatively and can form a co-functional module. This co-functional module is related to both of multi-cancers and multi-non-cancer diseases.

To reveal the detailed regulation mode of these miRNAs associating with multiple cancers and non-cancer diseases, we conducted a deep case analysis. In Figure 4.4, the top ten common target genes of each co-functional pair were combined to be a gene set. The DAVID functional annotation tool (Huang et al. 2009*b*, Huang et al. 2009*a*) was applied to analyze these gene sets of the co-functional pairs in the module miR-29a-miR-29b-miR-29c, where the threshold of the pathway enrichment analysis (Kanehisa & Goto 2000) was set as p-value$\leq$0.05. The labels on the edges from the diseases to the genes are the probabilities of genes to be the co-functional targets of the miRNA co-function module. For example, the edge from the diseases to the gene VEGFA has the label of "C 77% N 23%". This label means that the co-function module may dysregulate the gene VEGFA to contribute to the development of the 26 cancers (C) with the probability of 77%. This gene may also be the common target of the co-functional module during the dysregulation in the development of those 13 non-cancer disease (N) with the

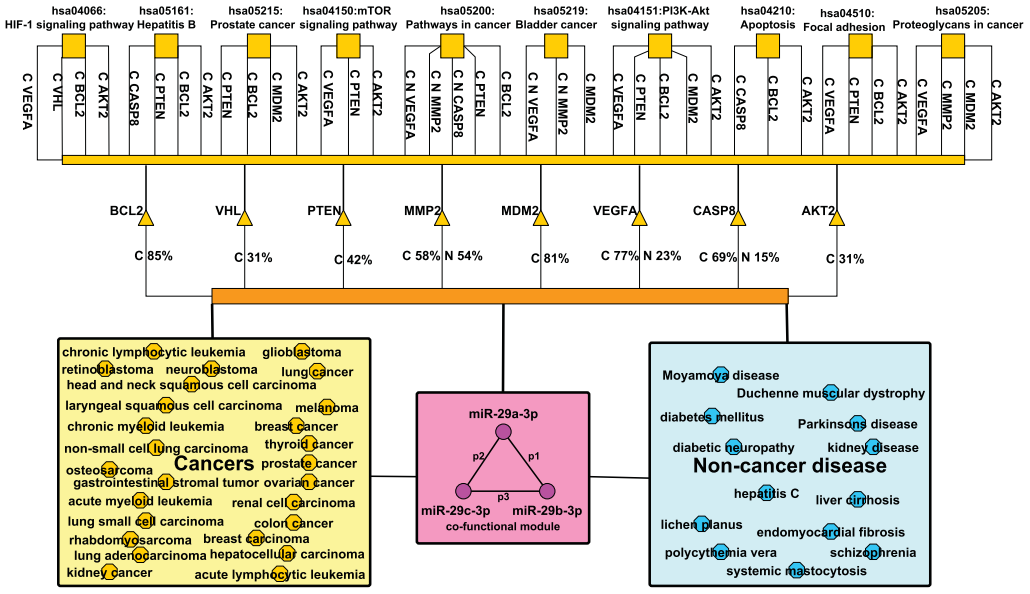


Figure 4.4: **The miR-29a-miR-29b-miR-29c co-function module, their targets and the enrichment analysis of the KEGG pathways.** The triangles are the potential common target genes of the miR-29a/b/c co-functional module. Those small squares are the genes enriched pathways. Those disease names in the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

probability of 23%. The labels along with the edges connecting the genes and the pathways indicate that the genes from the target gene sets of the diseases (i.e., cancers (C) or non-cancer diseases (N)) associated co-function module can be mapped to the corresponding pathways. For instance, there are three edges connecting the genes with the pathway 'hsa05219: Bladder cancer' together with the labels of "C N VEGFA", "C N MMP2" and "C MDM2". The labels mean the genes VEGFA, MMP2 and MDM2 from the target gene set of the cancers (C) associated co-function module can be mapped to the Bladder cancer pathway. For the non-cancer diseases (N), only two genes (VEGFA and MMP2) can be mapped to this pathway. Those genes that cannot map to any pathways or those diseases that are not associated with

all of the three co-functional pairs are ignored in the figure. The cancer related gene sets can be mapped to many different pathways, we just show the top ten pathways according to their p-values.

It is uncovered that the co-functional module mainly dysregulates the 'hsa05219: Bladder cancer' and the 'hsa05200: Pathways in cancer' to contribute to the development of the 13 non-cancer diseases. The module also regulates eight other pathways (hsa05205, hsa04510, hsa04066, hsa04151, hsa04150, hsa04210, hsa05161 and hsa05215) to involve in the development of the 26 cancers. The cancer developments are more complex with more common genes involved. This observation is consistent with the hypothesis that similar diseases may be related to similar miRNAs and genes. The top three non-cancer disease genes regulated by the co-functional module and mapped to the pathways are MMP2, VEGFA and CASP8, while for the cancers are BCL2, MDM2 and VEGFA. With the gene ontology enrichment analysis, we found that the former three genes have the function of angiogenesis (GO:0001525, p-value=1.8E-4), macrophage differentiation (GO:0030225, p-value=2.1E-3), negative regulation of cysteine-type endopeptidase activity involved in apoptotic process (GO:0043154, p-value=9.0E-3) and response to hypoxia(GO:0001666, p-value=2.2E-2). The latter three genes can play roles of cellular response to hypoxia (GO:0071456, p-value=5.2E-5), response to iron ion (GO:0010039, p-value=2.4E-3), ovarian follicle development (GO:0001541, p-value=5.8E-3) and the other related functions. The co-functional module can regulate two same pathways during the development of both the cancers and non-cancer diseases. The possible common targets also have the similar function such as response to hypoxia. These indicate that the miR-29a/b/c regulation module may contribute to the disease development partly via similar dysregulation mechanism. On the other side, the co-functional module may prefer to function by dysregulating the same genes in the development of various cancers rather than those non-cancer diseases. During the carcinogenesis of 26 kinds of cancers, averagely more than 70% of those cancers relate to the dysfunction of the above three

genes (BCL2, MDM2 and VEGFA). For the three non-cancer diseases related genes (MMP2, VEGFA and CASP8), the percentage is just around 30%. Those cancers related genes are more likely to involve in the same pathways which indicates the close relationships between their functions. This is mainly due to the fact that cancers are more similar to each other than those non-cancer diseases.

Interestingly, there are a number of literature which have reported the co-function of the miR-29 family members in the development of the cancers such as non-small-cell lung cancer (Tan, Wu & Cai 2013), renal cell carcinoma (Yonezawa, Enokida, Yoshino, Hidaka, Yamasaki, Itesako, Seki & Nakagawa 2013), breast cancer (Cittelly, Finlay-Schultz, Howe, Spoelstra, Axlund, Hendricks, Jacobsen, Sartorius & Richer 2013), ovarian cancer (Yu, Yan, Lai, Huang, Chou, Lin, Yeh & Lin 2014) and others types of cancers (Jiang, Zhang, Wu & Jiang 2014). Furthermore, the MYC-mediated miR-29 repression mechanism for the therapy of aggressive B-cell malignancies (B-cell malignancies is the synonym of chronic cymphocytic ceukemia according to Medical Subject Headings (MeSH) (Lipscomb 2000)) by applying the HDAC3 and EZH2 as therapeutic targets (Zhang, Zhao, Fiskus, Lin, Lwin, Rao, Zhang, Chan, Fu, Marquez et al. 2012) was reported. Another report in 2015 also discussed the adoption of miR-29s (miR-29a/b/c) as candidate epi-therapeutics for curing hematologic malignancies (Amodio, Rossi, Raimondi, Pitari, Botta, Tagliaferri & Tassone 2015). According to our findings and those articles, we claim that it is reasonable to consider miR-29a/b/c as potential drug targets for the treatment of multiple cancers.

Our conclusion is that the newly predicted disease related miRNAs and the prioritization of multi-disease associated co-functional miRNA pairs are highly effective for the analysis of the regulation mechanisms of miRNAs for different diseases at a system level. Particularly, it is useful to find common and special mechanisms in the development of different diseases and can provide new strategies for the diagnosis or treatment of the diseases. For example, if the three miRNAs (miR-29a/b/c) are proved to be effective drug

targets to cure some of the 26 cancers, they may also be suitable drug targets for the remaining cancers.

## 4.4 Conclusion

In this chapter, we proposed a novel method for prioritizing multi-disease associated co-functional miRNA pairs. It is the extended study of the research question **Q1** (see Section **1.2**). The contributions of this chapter's work are as follows (corresponding to the research contribution **C2** in Section **1.3**): (1) We reconstructed a disease-gene-miRNA tripartite network with our designed disease-miRNA association prediction method. The testing results show it provides more complete information for investigating the miRNA co-functioning; (2) We designed a scoring function to prioritize the candidate multi-disease associated co-functional miRNA pairs and their potential co-regulated genes; (3) We performed detailed case studies to understand the miRNA co-functional phenomenon for both cancers and non-cancer diseases; (4) We found that the multi-disease associated co-functional miRNA pairs can do good to the designing of multi-propose drugs for their related multi-diseases.

# Chapter 5

# Chromosome Preference of Disease Genes and Vectorization for the Prediction of Non-coding Disease Genes

## 5.1 Introduction

The background knowledge in Chapter 1 (Section **1.1**) shows that lncRNAs have been found to contain significant genetic information and functions (Mattick & Makunin 2006). The dysregulation of lncRNAs can result in the dysfunction of their target protein coding genes or their participated cellular processes, causing the development of diseases. Increasing number of studies have been focusing on the application of disease-lncRNA associations including disease diagnosis (Sánchez & Huarte 2013), survival prediction (Kumarswamy, Bauters, Volkmann, Maury, Fetisch, Holzmann, Lemesle, de Groote, Pinet & Thum 2014) and RNA therapeutics (Wahlestedt 2013). However, the function annotation of lncRNA genes such as their roles in disease development is remaining largely unknown.

Genomic locus inferring methods (Chen et al. 2013, Li, Gao, Wang,

Ma, Tu, Wang, Chen, Kong & Cui 2014), computational methods including gene-lncRNA co-expression methods (Sun, Luo, Liao, Bu, Zhao, Liu, Liu & Zhao 2013, Liu, Chen, Chen, Cui & Yan 2014), network analysis methods (Ganegoda et al. 2015), similarities analysis or semi-supervised learning methods (Chen & Yan 2013), supervised learning methods (Zhao, Xu, Liu, Bai, Xu, Xiao, Li & Zhang 2015) and others (Wang et al. 2016) can speed up this area of research for disease gene prediction. As concluded in Chapter 2, the network analysis heavily relies on the topology properties of the constructed networks. The semi-supervised learning methods depend on accurate similarity measurements between diseases and lncRNAs. The supervised learning approach has not been extensively explored because of lacking reliable negative samples of disease related lncRNA genes.

We propose to use a positive-unlabeled learning (PU-learning) method to predict disease related lncRNA genes. PU learning can well address the problem of lacking reliable negative samples to gain high prediction performance. In this work, we also introduce a novel vector $\langle V_d \rangle$ to represent a disease $d$, and a novel vector $\langle V_{Lnc} \rangle$ to represent an lncRNA gene *Lnc*. We merge these two vectors as $\langle V_d, V_{Lnc} \rangle$ to represent the pair of disease $d$ and the lncRNA gene *Lnc*. The prediction problem is: whether this merged vector can be mapped to 1 or 0 with a certain level of probability. If it is mapped to 1 with a high probability (e.g. 90%), then it means that the disease $d$ is related to the lncRNA gene *Lnc* under a high probability. Otherwise, the disease $d$ has little relationship with lncRNA gene *Lnc*.

The novel disease vector representation $\langle V_d \rangle$ consists of two sub-vectors. The elements of the first sub-vector $\langle V_d^{chr} \rangle$ represent the chromosome substructures' distribution information entropies of the genes related to the disease $d$. We consider 45 chromosome substructures in this work (details presented later).

This idea for disease representation is inspired by a chromosome substructure enrichment analysis of the disease related protein coding genes. It is similar to gene pathway enrichment analysis that the protein gene set

of a disease can be enriched at each chromosome substructure containing the protein gene set. We have observed that about 16.2% of 2802 diseases' genes can be enriched to chromosome 6 p-arm (with Fisher's exact test, p-value<0.05), implying a strong chromosome preference of disease genes. This preference is significantly higher than the second most enriched chromosome 2 q-arm (containing just 5.92% of the 2802 diseases). Furthermore, no disease gene set can be enriched to the chromosome 21 p-arm. Our hypothesis is that genes are located at various positions on chromosomes and mitochondrion, and the distribution of disease related protein coding genes on the chromosomes can be used to characterize the differences between diseases.

The second sub-vector $\left\langle V_d^{path} \right\rangle$ represents the KEGG pathway groups' distribution information entropies of disease $d$ related genes enriched KEGG pathways. Human KEGG pathways (Kanehisa & Goto 2000) can be divided into 30 groups. By the disease gene KEGG pathway enrichment analysis on the 2802 diseases, we have observed that almost all these KEGG pathways are involved in disease developments. The distribution of disease gene sets on KEGG pathway groups is also uneven. For example, more than 30% of the 2802 diseases are associated with 6 pathways including hsa04933: AGE-RAGE signaling pathway in diabetic complications and hsa05321: Inflammatory bowel disease (IBD). In comparison, as many as 61 kinds of pathways are related to less than 1% of these diseases.

Comparing with existing disease characterization methods through computing similarities of disease related coding or non-coding genes (Cheng et al. 2014), semantics (Mathur & Dinakarpandian 2012), phenotypes (Freudenberg & Propping 2002, Hoehndorf, Schofield & Gkoutos 2015), symptoms (Zhou, Menche, Barabási & Sharma 2014) and ontology (Li, Gong, Chen, Liu, Wu, Zhang, Li, Li, Rao & Li 2011), our disease vectorization $\left\langle V_d^{chr}, V_d^{path} \right\rangle$ is much simpler. It does not need repeated set operations such as union and intersection or large scale of text mining. Our disease vectors are also effective to capture unique disease characteristics. The disease similarity can reach

the average area under ROC curve (AUC) of 0.9458 when the diseases are represented by our vectors. However, FunSim (Cheng et al. 2014) and a disease symptom representation method (Zhou et al. 2014) have only 0.9202 and 0.7674 AUC respectively on the same set of diseases.

The vector $\langle V_{Lnc} \rangle$ representing an lncRNA gene *Lnc* consists of two sub-vectors $\langle V_{Lnc}^{seq} \rangle$ and $\left\langle V_{Lnc}^{prof} \right\rangle$ as well. The first one represents its sequence's k-mer frequencies, and the second one represents its expression profiles. Merging the two disease sub-vectors $\left\langle V_d^{chr} \right\rangle$ and $\left\langle V_d^{path} \right\rangle$, the two lncRNA sub-vectors $\langle V_{Lnc}^{seq} \rangle$ and $\left\langle V_{Lnc}^{prof} \right\rangle$, we can represent a disease-lncRNA gene pair (denoted $d - Lnc$) as $\langle V_d, V_{Lnc} \rangle$. Procedures for constructing the main sub-vectors are shown in Figure 5.1.

Disease related lncRNA genes should also prefer to co-expressing with other genes that are associated with this disease (such as those lncRNA genes which regulate some of the disease related protein coding genes). With this hypothesis, we add these co-expression features as the fifth sub-vector $\langle V_{co-exp} \rangle$ to the merged vector $\langle V_d, V_{Lnc} \rangle$. From our baseline classifier selection experiments, we have proved that this new sub-vector can further improve the prediction performance.

A bagging SVM for PU learning algorithm (Mordelet & Vert 2014) is adopted to prioritize disease related lncRNA genes. This model was trained on a set of disease-lncRNA vectors. On three data sets retrieved from three disease-lncRNA association databases: LncRNADisease (Chen et al. 2013), Lnc2Cancer (Ning et al. 2016) and MNDR (Wang, Chen, Chen, Li, Kang, Fan, Hu, Xu, Yi, Yang et al. 2013), the overall AUC scores of leave-one-out cross-validation (LOOCV) by our method are 0.8016, 0.8335 and 0.7527 respectively. This performance is significantly superior to two state-of-the-art methods: LRLSLDA (Chen & Yan 2013) (0.6882, 0.7308 and 0.6346) and LRLSLDA-ILNCSIM (Huang, Chen, You, Huang & Chan 2016) (0.6949, 0.7390 and 0.6435). Especially when only the sequence information of the lncRNA genes is available, our method can still work well for the prediction. The overall LOOCV AUC scores for the three datasets are 0.7889, 0.8266 and

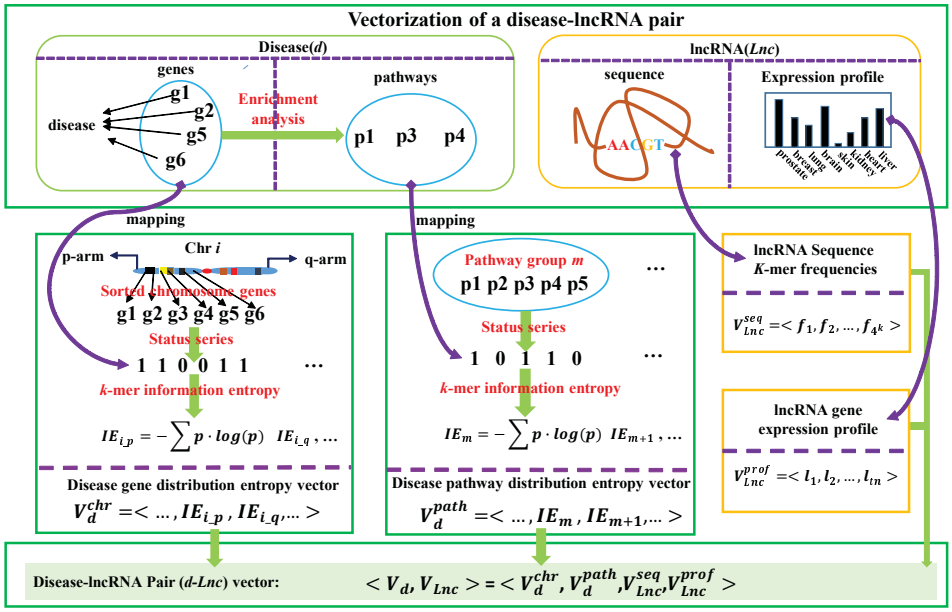


Figure 5.1: **The flowchart for the vectorization representation of a disease-lncRNA gene pair.** A disease-lncRNA gene pair can be represented by the integration of four sub-vectors including disease gene chromosome substructures' distribution information entropy vector (disease gene distribution vector), 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.

0.7216. The results of the following leave-one-disease-out cross-validation (LODOCV) experiments show the ability of our method to predict without known disease related lncRNA genes for a given disease as the average AUC value is 0.7356 for the LncRNADisease dataset. There are 68 out of 162 diseases can achieve the AUC values bigger than 0.9.

## 5.2 Materials and Methods

Datasets of diseases and disease related genes were collected and relevant human KEGG pathways were collected as well for the construction of the disease vectorization model and the disease gene prediction method. The details of the datasets and prediction algorithms are presented below.

### 5.2.1 Diseases, disease genes and KEGG pathways

The Medical Subject Headings (MeSH) (Lipscomb 2000), Comparative Toxicogenomics Database (CTD) (Davis et al. 2009), Disease Ontology (DO) (Schriml et al. 2012) and Online Mendelian Inheritance in Man (OMIM) (Hamosh et al. 2005) are widely visited databases containing massive amount of disease related information. However, there is no standard for the adoption of disease names or ids between these databases. We mapped disease names to DO ids using the DO, MeSH and CTD as dictionaries. Similarly, for genes, we did id or name conversion using the data records from the HUGO Gene Nomenclature Committee (HGNC) (Povey et al. 2001) database. It contains reference records of genes among a great number of widely used databases. In this work, we mainly mapped the genes obtained from various resources to entrez gene ids (Maglott, Ostell, Pruitt & Tatusova 2005). We downloaded the HGNC database on Jun 17, 2016. There are totally 39670 approved gene records with entrez gene ids including 19025 protein coding genes and 20645 non-protein coding genes.

We downloaded disease-gene associations from the supplementary file of a published article (Cheng et al. 2014) which contains 117,190 associations between 2817 diseases and 12063 genes. The authors collected these data records from database SIDD (Cheng et al. 2013). Each of the diseases has a unique id from database DO. After data correction and redundancy removal according to the latest version of the databases DO, MeSH, CTD and HGNC, we obtained a set of 114754 disease-gene associations between 2802 diseases and 10893 genes (including 10321 protein coding genes and 572 non-protein

coding genes). The human KEGG pathways were extracted from the KEGG database on June 21, 2016. There are 303 unique pathways containing 7060 unique genes (all have an entrez gene id). All these datasets are listed in Supplementary file 9 and Supplementary file 10.

### 5.2.2 Associations between diseases and lncRNAs

The disease-lncRNA associations were obtained from three databases: lncRNADisease (downloaded on April 18, 2016), lnc2cancer (downloaded on July 4, 2016) and MNDR (downloaded on June 30, 2016). There are 1102, 1239 and 754 disease-lncRNA associations (redundant and unclear information are existing). For the diseases, we mapped them to DO. To construct our PU learning model for disease related lncRNA prediction, we collected the sequences and expression profiles of the lncRNAs.

We mapped each of these lncRNAs to its corresponding ensembl gene id, RefSeq accession id, entrez gene id and other detail information. This process was manually finished via searching and comparing the lncRNA related databases such as ensembl (Cunningham, Amode, Barrell, Beal, Billis, Brent, Carvalho-Silva, Clapham, Coates, Fitzgerald et al. 2014), NONCODE (Zhao, Li, Fang, Kang, Wu, Hao, Li, Bu, Sun, Zhang et al. 2015), Lncipedia (Volders, Helsens, Wang, Menten, Martens, Gevaert, Vandesompele & Mestdagh 2012), lncRNAdb (Quek, Thomson, Maag, Bartonicek, Signal, Clark, Gloss & Dinger 2014), HGNC. Then, lncRNA sequences were extracted from the RefSeq (Pruitt, Tatusova & Maglott 2006). We downloaded the expression level of 60245 genes (coding or non-coding genes matched with an ensembl id and gene symbol) in 16 tissues from the Expression Atlas (Petryszak, Keays, Tang, Fonseca, Barrera, Burdett, Füllgrabe, Fuentes, Jupp, Koskinen et al. 2015).

Finally, we obtained 454 disease-lncRNA associations from lncRNADisease (between 162 diseases having known disease genes and 187 lncRNAs with known sequences and expression levels). Those 594 (79 cancers, 310 lncRNAs) and 176 (86 diseases, 57 lncRNAs) more pairs were extracted

from lnc2cancer and MNDR respectively. For those diseases that not exist in the above 2802 ones, disease genes were obtained from the CTD, DisgeNet (Bauer-Mehren, Rautschka, Sanz & Furlong 2010), OMIM and malaCard (Rappaport, Nativ, Stelzer, Twik, Guan-Golan, Iny Stein, Bahir, Belinky, Morrey, Safran & Lancet 2013). The datasets are stored in Supplementary file 11.

### 5.2.3 Disease gene chromosome preference analysis and disease vectorization method

Human genes are located on mitochondrion and 24 unique chromosomes including 22 autosomes and two sex chromosomes. The genes' locations on the chromosomes or mitochondrion have been labeled by the HGNC database. As disease related genes are distributed at various locations and have a different number of each disease, we hypothesize that the gene distribution differences between diseases on the chromosomes or mitochondrion may reflect the divergences of the diseases. We also hypothesize that disease genes may have some preferred chromosomes for some diseases. This hypothesis can be investigated by the disease genes' chromosomes enrichment analysis via fisher's exact test (Beißbarth & Speed 2004). Thus, it is better to characterize the distribution properties of disease genes on each of the chromosomes instead of on the whole known gene set (we call it a "part overcomes the whole" hypothesis).

On the basis of these hypotheses, we considered to vectorize a disease via modeling the distribution property of its related gene set. However, with just the gene distribution characteristics, there may be no gene function information involved. Thus, we considered to extract the distribution properties of disease gene enriched KEGG pathways comparing to all the known pathways to inject complementary information for our vector representation of diseases. This vectorization process includes the following steps:

- Step1: Initialization. Sorting all known genes according to their chromosome locations and sorting all the human KEGG pathways by their ids.

- Step2: Grouping. Dividing the genes and pathways into groups. Producing a status series for each group with the length equals to the number of genes or pathways it contains. These statuses are initialized to be 0 (inactivated).

- Step3: Mapping. For a given disease related gene set, mapping them to the gene groups and mapping its enriched pathways to the pathway groups. Then, setting the corresponding status of a gene or pathway in the status series to be 1 (activated) if it has been mapped.

- Step4: Vectorization. Calculating the status series' k-mer information entropy of each gene group or pathway group to quantify them and constructing two sub-vectors for a given disease.

Here, dividing all the genes and pathways into groups is to apply our "part overcomes the whole" hypothesis. In our Results section, we demonstrate that this strategy (part) is more effective for characterizing diseases comparing to the status series without dividing (the whole). As a chromosome always contains a p-arm and a q-arm, we divide the genes into groups according to the natural chromosome substructures. For the pathway status series, we divide it into $T$ groups on average. (There is no guidance for us to divide pathways similar to chromosome structure). Finally, this vectorization model includes two parts: disease gene set vectorization and disease gene enriched pathway set vectorization.

Let $d$ represent a disease, and $d_g = \{g_1, g_2, \cdots, g_k, \cdots, g_n\}$ be its related gene set. Let all of the approved genes from HGNC be $G = \{g_1, g_2, \cdots, g_i, \cdots, g_N\}$, and the pathway set from KEGG database be $P = \{p_1, p_2, \cdots, p_j, \cdots, p_M\}$. Let the unique genes in $P$ be represented as $G_p = \{g_1, g_2, \cdots, g_A\}$ while each pathway related gene set as $g_{pj}$. We define

a k-mer sub-status series as $(s_1, s_2, \cdots, s_r, \cdots, s_k)$, where $s_r = 0$ or 1. By definition, there can be $2^k$ possible k-mer sub-status series. The detail process is described as a pseudo codes in **Algorithm** 5.1 and outlined in Figure 5.1. The source codes can be referred to Supplementary file 12.

Then we compute the similarity between $d1$ and $d2$ with their vectors. The similarity between $d1$ and $d2$ is denoted as $Sim(d1, d2)$ and computed by:

$$simGe\left(D_1, D_2\right) = \begin{cases} 0 \quad if\ ||E^{ge}_{D1}|| \times ||E^{ge}_{D2}|| = 0; \\ subspace\left(E^{ge}_{D1}, E^{ge}_{D2}\right) \quad else. \end{cases} \tag{5.1}$$

$$simPe\left(D_1, D_2\right) = \begin{cases} 0 \quad if\ ||E^{pe}_{D1}|| \times ||E^{pe}_{D2}|| = 0; \\ subspace\left(E^{pe}_{D1}, E^{pe}_{D2}\right) \quad else. \end{cases} \tag{5.2}$$

$$Sim\left(D_1, D_2\right) = e^{-[\theta \times simGe + (1-\theta) \times simPe]} \tag{5.3}$$

where $\theta$ is a parameter to mediate the ratio of each vector's contribution to compute the similarity. $||\cdot||$ means the norm. $subspace(x, y)$ is the function to obtain the angle between two vectors $x$ and $y$. Larger value of $Sim\left(D_1, D_2\right)$ shows more similarity of the diseases.

The four parameters $k1$ (the size of k-mer for gene series), $k2$ (the size of k-mer for pathway series), $T$ and $\theta$ can be determined via a performance test through comparing the disease similarity with a benchmark dataset. We first set $\theta$=1 to optimize $k1$ and set $\theta$=0 to optimize $k2$, $T$ with the objective of achieving the best performance. Then, the three parameters are set as the optimal values to select the best $\theta$. Similarly, we can also apply subspaces between the disease gene status series (a disease is represented as a fixed-length vector with the elements equal to 0 or 1) or the pathway status series themselves instead to measure the similarities of diseases. We call them the disease gene status series vector method and the pathway status series vector method. In the Results section, we compare the performances of our status series methods and our entropy vector methods to prove our "part overcomes the whole" hypothesis.

**Algorithm 5.1 Disease vectorization.**

**Require:** disease $d$ related gene set $d_g$, the number of genes in $d_g$ is $n$, Approved genes $G$, human pathway set $P$ with totally $M$ pathways, the number of unique genes in $P$ is $A$, each pathway $p_j$ contained gene set $g_{pj}$, parameter $k1$, $k2$, $T$;

1: Sort $G$ according to the chromosome location of $g_i$, sort $P$ according to the ids of $p_j$;
2: Separate $G$ according to the natural chromosome structure such as chr1 p-arm, chr1 q-arm, $\cdots$. There are total $S$ chromosome substructures, i.e. $chr_1, chr_2, \cdots, chr_u, \cdots, chr_S$;
3: Divide $P$ into $T$ groups, i.e. $p_{g1}, p_{g2}, \cdot, p_{gw}, \cdots, p_{gT}$;
4: Generate the initial status series of each $S\_chr_u = (0, 0, \cdots, 0)$ and $S\_pg_w = (0, 0, \cdots, 0)$;
5: Map $d_g$ to $chr_u$ according to its location and change the corresponding status in $S\_chr_u$ as 1;
6: Set $k$=$k1$
7: **for** $u = 1$ **to** $S$ **do**
8: Scan $S\_chr_u$ with window size of $k$ and step size 1;
9: Compute the frequency of $q$th $k$-mer sub-status series as $f_q$
10: Compute the summary of the entropy of all the $k$-mer sub-status series for $S\_chr_u$ as $E_{chr_u} = \sum_{q=1}^{2^k} f_q log\,(f_q)$ ;
11: **end for**
12: **for** $j = 1$ **to** $M$ **do**
13: Count genes in $g_{pj}$ as $L = Length\,(g_{pj})$;
14: Count genes mapped into $p_j$ as $B = Length\,(d_g \cap g_{pj})$;
15: Do the fisher exact test:
16: $[h, p, stats] = fishertest\,([L - B, B; A - L - n + B, n - B])$, where $p$ is the p-value;
17: **if** $p \leq 0.05$ **then**
18: Change the corresponding status of $p_j$ in $S\_pg_w$ as 1;
19: **end if**
20: **end for**
21: Set $k = k2$
22: **for** $w = 1$ **to** $T$ **do**
23: Scan $S\_pg_w$ with window size of $k$ and step size 1;
24: Compute the frequency of $v$th $k$-mer sub-status series as $f_v$
25: Compute the summary of the entropy of the all the $k$-mer sub-status series for $S\_pg_w$ as $E_{pg_w} = -\sum_{v=1}^{2^k} f_v log\,(f_v)$ ;
26: **end for**
27: **Output** The gene set entropy vector $E_D^g = [E_{chr_1}, E_{chr_2}, \cdots, E_{chr_S}]$
28: **Output** The genes enriched pathway entropy vector $E_D^p = [E_{pg_1}, E_{pg_2}, \cdots, E_{pg_T}]$

### 5.2.4 Prioritizing disease related lncRNA genes

We always just have the positive samples of disease-lncRNA associations, as the negative samples, namely the lncRNAs that do not relate to the diseases, are neglected or even cannot be obtained. Supervised learning algorithms are unable to deal with this situation. However, the Positive Unlabeled learning (PU learning) method (Li & Liu 2003) can address this issue effectively. PU learning has been an effective method for solving similar problems in bioinformatics such as disease gene prediction (Yang, Li, Mei, Kwoh & Ng 2012), predicting conformational B-cell epitopes (Ren, Liu, Ellis & Li 2015), splicing elucidation (Hao, Colak, Teyra, Corbi-Verge, Ignatchenko, Hahne, Wilhelm, Kuster, Braun, Kaida et al. 2015) and kinase substrate prediction (Yang, Humphrey, James, Yang & Jothi 2015). These PU learning approaches are mainly derived from two types of PU learning algorithms: the biased-SVM (Liu, Dai, Li, Lee & Yu 2003) and Elkan et al's lemmas (Elkan & Noto 2008). The application of Elkan et al's lemmas requires the satisfaction of "selected completely at random assumption", while the biased-SVM methods need to tune a set of parameters. Mordelet et al. (Mordelet & Vert 2014) proposed a bagging SVM model for PU learning and proved that their model can match and even outperform the biased-SVM algorithm. Especially, the bagging SVM for PU learning algorithm can run considerably faster. We adopt this bagging SVM PU learning to prioritize disease related lncRNA genes.

Let $Lnc$ be a lncRNA gene, represented as $Lnc{=}l_1 l_2 \dots l_e \dots l_O$. We calculate its k-mer frequency $\langle V_{Lnc}^{seq} \rangle$ and its expression profile $\left\langle V_{Lnc}^{prof} \right\rangle$. As there are four kinds of nucleotides in a lncRNA sequence (i.e., $l_e \in \{A, G, C, T\}$), there are $4^k$ possible k-mers. These k-mers are sorted by their alphabetic order. Their frequencies are counted via the window sliding mechanism with the window size of $k$ and a step size 1, which are then the elements of the vector $\langle V_{Lnc}^{seq} \rangle$. The expression profile of $Lnc$ can be extracted from the Expression Atlas (Petryszak et al. 2015). The expression levels of the lncRNA gene in the 16 tissues are the elements of the vector $\left\langle V_{Lnc}^{prof} \right\rangle$.

Then for a disease-lncRNA pair, e.g. $d - Lnc$, we construct another feature namely vector $\langle V_{co-exp} \rangle$, called the co-expression levels. This sub-vector can be constructed on the basis of the principle that a disease related lncRNA gene may show the preference of co-expressing with other genes associating with this disease (such as the lncRNA's targets). This sub-vector contains three elements, i.e. the maximum, minimum and average spearman correlation coefficients ($\langle V_{co-exp} \rangle = \langle max_{co-exp}, min_{co-exp}, avg_{co-exp} \rangle$) between the expression profile of $Lnc$ and all the known disease $d$ related genes' expression profiles.

The whole disease-lncRNA feature vector is formed by combining the five sub-vectors: the disease gene distribution entropy vector $\left\langle V_d^{chr} \right\rangle$ (sf1), disease pathway distribution entropy vector $\left\langle V_d^{path} \right\rangle$ (sf2), lncRNA sequence's k-mer frequency $\langle V_{Lnc}^{seq} \rangle$ (sf3), lncRNA expression profile $\left\langle V_{Lnc}^{prof} \right\rangle$ (sf4), and the co-expression features $\langle V_{co-exp} \rangle$ (sf5). The pseudo codes for prioritizing the disease related lncRNAs with the bagging SVM for PU learning model are shown as **Algorithm** 5.2.

In **Algorithm** 5.2, $|PO|$ means the sample size of the positive dataset. The feature type means the type of combination of the five sub-vectors. The feature vector $\left\langle V_d^{chr}, V_{Lnc}^{seq} \right\rangle$ is used as the basic feature type. Adding the remaining sub-features to this basic type makes new feature types. The best one can be identified via comparing the results of the cross-validation experiments. The bootstrap strategy is adopted with the purposes of making good use of the abundant unlabeled samples and improving the prediction performance. After obtaining the scores for the unlabeled samples, we sort them. The larger scores imply that the samples are more likely to be positive ones.

**Algorithm 5.2 A bagging SVM for prioritizing the disease related lncRNA genes.**

**Require:** Positive dataset $PO$, unlabeled dataset $UN$, bootstrap sample size $R$, bootstrap number $V$, SVM parameters, feature type $W$

1: **for** $a$=1 **to** 100 **do**
2: Randomly select $|PO|$ of unlabeled samples as negative samples
3: Implement a 5-fold cross validation on the positive-negative dataset with feature type $W$ and do grid search of SVM parameters;
4: **end for**
5: Use the $F1$ score as the metric, determine the optimal SVM parameters $opPara$ and the optimal feature type $W_{op}$;
6: $\forall x \in UN, n(x) \leftarrow 0, f(x) \leftarrow 0$;
7: **for** $b$=1 **to** $V$ **do**
8: Draw a bootstrap sample $UN_b$ of size $R$ in $UN$
9: Train a classifier $f_b$ to discriminate $PO$ against $UN_b$ with $opPara$ and $W_{op}$;
10: For any $x \in UN \setminus UN_b$, update:
11: $f(x) \leftarrow f(x) + f_b(x)$;
12: $n(x) \leftarrow n(x) + 1$.
13: **end for**
14: **Output** The score $s(x) = f(x)/n(x)$ for $x \in UN$.

## 5.3 Results

### 5.3.1 Chromosome preference and disfavor of disease genes

In the understanding of unique characteristics of disease genes on the chromosomes, we constructed chromosome enrichment analysis of disease genes. The process is similar to the implementation of Fisher's exact test for pathway enrichment analysis which we have described in **Algorithm** 5.1. The main difference is that the pathway genes are replaced with the chromosome involved genes. We note that only protein coding genes are considered for the chromosome preference analysis of disease genes as the non-coding disease genes are under prediction.

The 24 chromosomes of human genome can be naturally divided into

48 substructures with the p-arm and the q-arm as two substructures for each chromosome. However, for chromosome 13 (*chr*13), there is only one protein gene on the centromere and there is no approved protein gene located at its p-arm; for chromosome 14, only one gene is located at its p-arm; and there is no gene located at the p-arm of *chr*15 or *chr*22. Thus, these four chromosomes were not divided. We consider the mitochondrion as a special chromosome which cannot be divided into two substructures. In total, we have 45 chromosome substructures, namely $S$=45 in **Algorithm** 5.1. Figure 5.2 and Figure 5.3 show the statistics of the chromosome substructure enrichment analysis for the disease genes of each of the 2802 diseases.

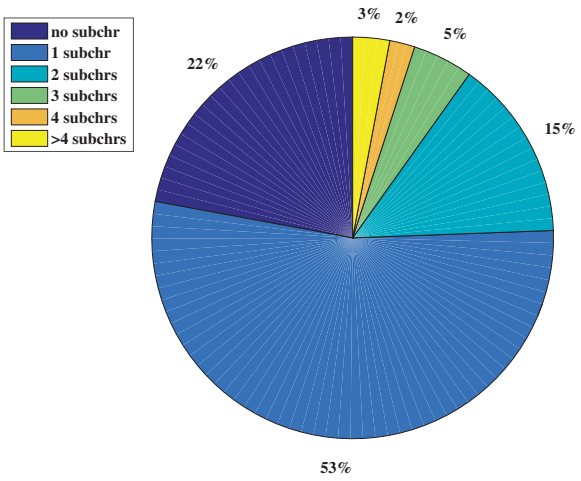


Figure 5.2: **The disease chromosome enrichment analysis pie graph.** Subchr means chromosome substructure. We did the statistics of how many chromosomes a disease gene set enriches. More than a half (53%) of the 2802 diseases are just enriched to only one chromosome substructure, while just 3% of these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

There are about 75% of the diseases whose related gene sets can be enriched to no more than 1 chromosome substructure (Figure 5.2). There

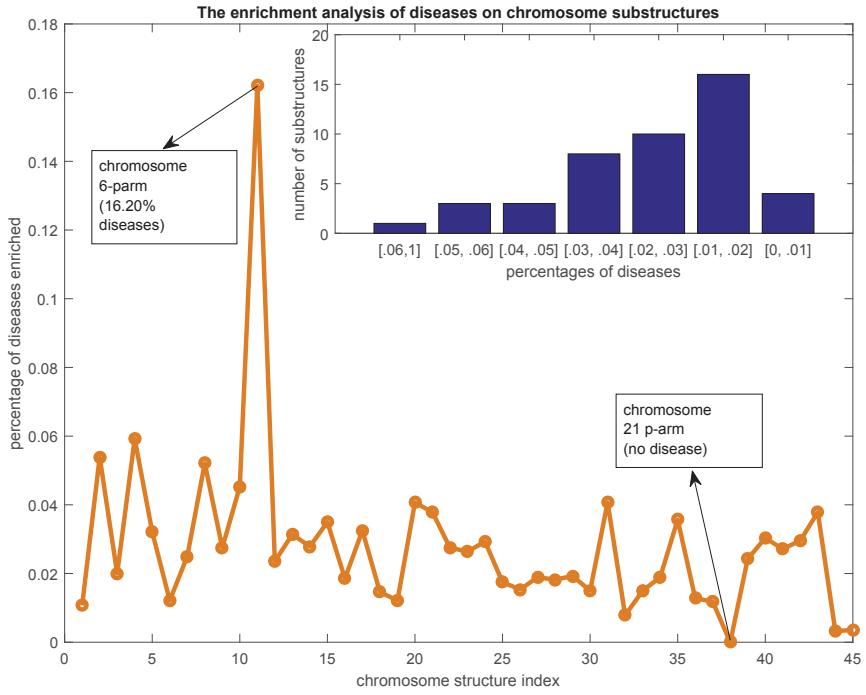


Figure 5.3: **The disease chromosome enrichment analysis results.** The y-axis are percentages of diseases that enriched to each of the chromosome substructures. The x-axis are the indexes of the chromosome substructures. The bar graph at the top right shows the statistics of the numbers of chromosome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

are just 3% of the diseases whose related gene sets can be enriched to more than 4 chromosome substructures. This distribution of disease genes on the chromosome substructures reveals that the disease genes of a given disease are very likely located at a neighborhood region. As indicated by Figure 5.3, the p-arm of chromosome 6 is the most preferred substructure of disease genes — about 16.2% of the disease related gene sets can be enriched here. This percentage is significantly higher than the other substructures

(all no more than 6%). Interestingly, there is no disease related gene set that can be enriched to the p-arm of chromosome 21. From the top-right bar graph of Figure 5.3, we can also see that 16 out of the 45 chromosome substructures are enriched by only 1%-2% of the 2802 gene sets. There are 10 and 8 substructures can be enriched by 2%-3% and 3%-4% of the 2802 gene sets. Thus, most of the chromosome substructures (38 out of 45) can be enriched by no more than 3% of the 2802 gene sets. These observations suggest a phenomenon that disease genes are unevenly distributed in the 45 chromosome substructures. The genes related to a disease are preferred at a physical neighborhood close to each other in the chromosomes. This observation of chromosome preference lays down the foundation for our disease vector representation.

We also conducted pathway enrichment analysis to understand the distribution of disease genes in human KEGG pathways. We found that disease genes are also unevenly enriched in these pathways. More than 30% of the 2802 diseases are associated with one of the top 6 pathways such as hsa04933: AGE-RAGE signaling pathway in diabetic complications, and hsa05321: Inflammatory bowel disease (IBD). In contrast, 61 out of 303 pathways are related to less than 1% of these diseases. More details are reported in Supplementary file 13.

### 5.3.2 Performance on the prediction of highly similar diseases using our disease vector representation

We tested the performance of our vectorization model for computing disease similarities on the dataset downloaded from the supplementary files of Cheng's paper (Cheng et al. 2014). It contains a candidate disease set and a benchmark set of similar disease pairs. The disease set is composed of 2802 diseases and their related genes. There are 70 similar disease pairs in the benchmark set. Zhou et al. (Zhou et al. 2014) proposed a symptom representation method for measuring disease similarities. To compare this method with ours, we downloaded their similarity scores between 1596

diseases and mapped these diseases to the disease set. Totally 1012 diseases and 56 similar disease pairs in the benchmark set can be mapped. These two disease sets have been stored in Supplementary file 14.

Following cheng's method, we drew a ROC curve to display how our method can rank the similar pairs in the benchmark set comparing with those randomly selected unknown disease pairs. That means, for a given threshold, if the similarity of a pair in the benchmark set exceeds this threshold, it is defined as a true positive, otherwise, as a false negative. Inversely, an unknown disease pair exceeds the threshold is defined as a false positive. A total of 560 testing disease-disease pairs were randomly selected from the 1012 candidate diseases (but not overlapping with the benchmark set). This process was repeated 100 times.

There are three parameters, i.e. $k1$, $k2$ and $T$, for **Algorithm** 5.1 and one parameter $\theta$ for equation 5.3 need to be tuned. According to the HGNC database, there are 19025 approved protein coding genes. Because the minimum length of the chromosome substructure is 9 (only 9 protein coding genes on this substructure), thus the parameters $k1$ was changed from 1 to 9 with the step size of 1. There are 303 different human KEGG pathways. To simplify our model, we set $T$=30 with the first 29 groups containing 10 pathways while the last group has 13 pathways. Finally, the disease genes chromosome substructures' distribution information entropy (disease gene distribution entropy) feature is represented as a 45-dimensional vector while the disease gene enriched pathway groups' distribution information entropy (disease pathway distribution entropy) feature is a 30-dimensional vector. $k2$ is changed from 1 to 10 with the step size of 1. The integration parameter $\theta$ is in the range of [0, 1].

When $k1$=9, we can get the biggest average AUC=0.9429. Meanwhile, when $k2$=8, the AUC value with just pathway distribution entropy vectors can achieve 0.8872. Thus, we set $k1$=9 and $k2$=8 for the subsequent experiments.

We also compared the performances of our methods (namely the entropy

vector methods and the status series vector methods), the FunSim (Cheng et al. 2014) and symptom representation method (Zhou et al. 2014). We implemented the FunSim according to the published paper. Then, the AUC values were computed according to the scores via different methods. During the comparison, $\theta$ was set to be 0 to 1 with the step size of 0.1. When $\theta$=0.8, the integrated similarity method can work best with average AUC=0.9458. We drew the corresponding overall ROC curves (all the 100 times repeat experiments' results are combined together to compute the False Positive Rate and True Positive Rate; thus, the overall AUC values are smaller than the average AUC values) of the 100 times experiments in Figure 5.4. More comparison results for the original 2802 disease set can be found in the Supplementary file 13.

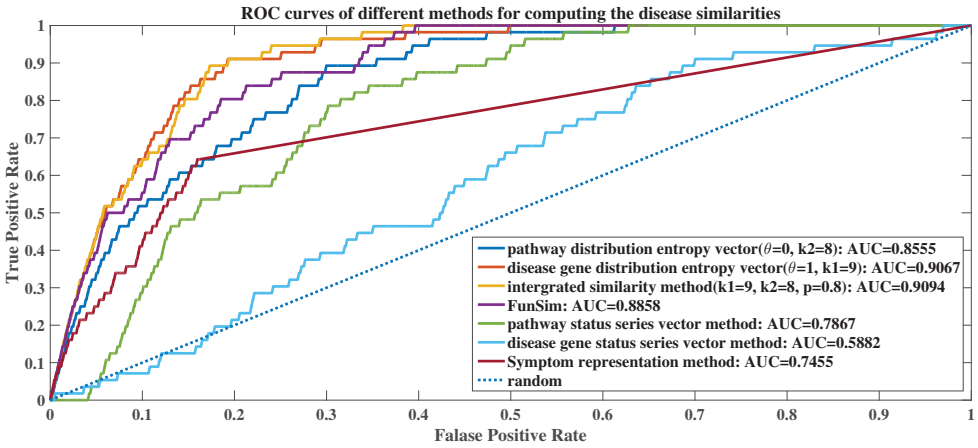


Figure 5.4: **The ROC curves of different methods for computing the disease similarities.** There are 7 ROC curves: the disease pathway distribution entropy vector method ($\theta$=0, AUC=0.8555); the disease gene distribution entropy vector method ($\theta$=1, AUC=0.9067); the integrated similarity method ($\theta$=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).

Figure 5.4 shows that the integrated similarity method is better than the other methods. However, it just improves 0.0027 on the AUC value comparing with just disease gene distribution entropy vector method ($k1$=9, $\theta$=1). It implies that there is not much complementary between the disease gene distribution entropy and disease pathway distribution entropy features as to compute the similarities of diseases. The integrated similarity method with $\theta$=0.8 outperforms the FunSim and symptom representation method by improving AUC values of 0.0236 and 0.1639 respectively. In comparison, the status series vector methods cannot work as well as the entropy vector methods. The entropy vector methods (disease gene distribution entropy vs. disease gene status series and disease pathway distribution entropy vs. pathway status series) improve the overall AUC values by 0.3185 and 0.0688. This proves our "part overcomes the whole" hypothesis that our dividing and information entropy strategy for representing diseases is more effective than the original status series.

### 5.3.3 Performance on the prediction and prioritization of disease related lncRNA genes

The performance of our disease vectorization method for predicting and prioritizing disease related lncRNA genes was tested and evaluated on three datasets: the lncRNADisease dataset (454 positive samples, i.e., 454 known associations between some diseases and some lncRNA genes), the lnc2cancer dataset (594 positive samples) and the MNDR dataset (176 positive samples). See details of these datasets at the section **Materials and Methods**.

#### Classifier and parameter selection for final prediction model with the *lncRNADisease dataset*

We used both liner and RBF kernel for the SVM-based positive-unlabled learning method to conduct cross-validation on the lncRNADisease dataset. The number of positive samples is 454, and the number of unlabeled samples

Table 5.1: **Feature types and their corresponding performance.**

| Type | Combinations | Liner kernel | RBF kernel |
|---|---|---|---|
| 0 | sf1, sf3 | C = 7, F1 =0.6668 | C = -1, G = -1, F1 = 0.6734 |
| 1 | sf1, sf2, sf3 | C = 2, F1 =0.6895 | C = 3, G = -5, F1= 0.7024 |
| 2 | sf1, sf3, sf4 | C = 7, F1 =0.6692 | C = 6, G = -2, F1 = 0.6734 |
| 3 | sf1, sf2, sf3, sf4 | C = 0, F1 =0.6942 | C = 5, G = -7, F1 = 0.7058 |
| 4 | sf1, sf3, sf5 | C = 8, F1 =0.6658 | C = 0, G = -2, F1 = 0.6768 |
| 5 | sf1, sf2, sf3, sf5 | C = 0, F1 =0.6906 | C = 4, G = -6, F1 =0.7032 |
| 6 | sf1, sf3, sf4, sf5 | C = 1, F1 =0.6708 | C = 0, G = -2, F1 = 0.6748 |
| 7 | sf1, sf2, sf3, sf4, sf5 | C = 2, F1 =0.7004 | C = 3, G = -5, F1 = 0.7114 |

(i.e., the number of unknown associations) is 29840, derived by exhaustively pairing the 162 diseases and 187 lncRNAs in the lncRNADisease dataset after the deduction of the number of 454 positive samples. Recall that our vector representation for a pair of disease and lncRNA gene consists of five sub-vectors. Here, we choose different combinations of these sub-vectors to understand that all of these sub-vectors are important for the prediction. The steps are presented in **Algorithm** 5.2.

The basic combination of the sub-vectors is to merge the disease gene distribution entropy sub-vector $\langle V_d^{chr} \rangle$, and lncRNA sequence's k-mer frequency sub-vector $\langle V_{Lnc}^{seq} \rangle$. Here, we set $k$=3 (k-mer size for lncRNA sequence) and $k1$=9 (k-mer size for disease gene series) in the previous section. This basic feature vector is a 109-dimensional (45+64) feature vector, simply denoted by sf1+sf3. We name it the type-0 feature vector. Adding other sub-vectors such as the disease pathway distribution entropy vector $\left\langle V_d^{path} \right\rangle$ (sf2, 30-dimensional), lncRNA expression profile $\left\langle V_{Lnc}^{prof} \right\rangle$ (sf4, 16-dimensional), the basic feature vector can be expanded into another three feature types, i.e., the feature type 1-3 in Table 5.1. Furthermore, the co-expression feature namely the fifth sub-vector $\langle V_{co-exp} \rangle$ (sf5, 3-dimensional) was added to each of the former combinations and form four more feature types which are showed in the last four lines of Table 5.1.

Adding the disease pathway distribution entropy sub-vector $\left\langle V_d^{path} \right\rangle$ (i.e., sf2) can improve the performance for predicting disease-lncRNA associations

(type1 vs. type0, type3 vs. type2, type7 vs. type6, averagely improved by 0.0257 for liner SVM and 0.0307 for RBF SVM respectively). However, the improvement by adding the lncRNA expression profile is not as high as adding the disease pathway distribution entropy sub-vector (0.0052 for liner SVM, 0.0021 for RBF SVM averagely). However, the co-expression feature vector $\langle V_{co-exp} \rangle$ can further improve the prediction performance averagely by 0.0039 and 0.0033 for liner SVM and RBF SVM respectively. The combination of all the 5 sub-vectors (i.e., the type 7 feature vector) worked the best among the 8 types of feature vectors (on average improving by 0.0223 for liner SVM and 0.0243 for RBF SVM). Furthermore, the RBF kernel outperforms the liner kernel (on average improving by 0.0092). Thus, our baseline classifier is the RBF SVM ($C$ = 3, $G$ = -5) with the type 7 feature vector representation ($W$=7).

Using all the sub-vectors (i.e., the type 7 feature vector) to represent a pair of disease and lncRNA gene, the 5-fold cross-validation AUC results on the lncRNADisease dataset by bagging SVM is showed in Figure 5.5, using different bootstrap sample size $R$ and the bootstrap number $V$. Here, we repeated the experiment 10 times. The AUC values were computed by comparing the scores of known pairs (set to be unknown during the cross validation) with those unknown ones. We note that we simply set $R = |PO|$ as Mordelet et al (Mordelet & Vert 2014) had proved that setting $R$ to be the same as the size of positive samples is a safe choice for the bagging SVM.

The AUC values change in a narrow scope (0.79-0.81) when the bootstrap number $V$ varies from 10 to 400. In fact, the running time for computing the scores of unknown samples increases significantly when $V$ is increasing. As bigger $V$ achieves weak improvement of the performance but results in significant increase of time cost, we suggest fixing $V$=10. This is consistent with the conclusion of Mordelet's report that when $R$ is large, the SVM usually rarely benefits from bagging. Thus, our final PU learning classifier is built with following parameters: RBF kernel SVM with $C$ = 3, $G$ = -5, $V$ = 10, $R = |PO|$ and $W$ = 7.

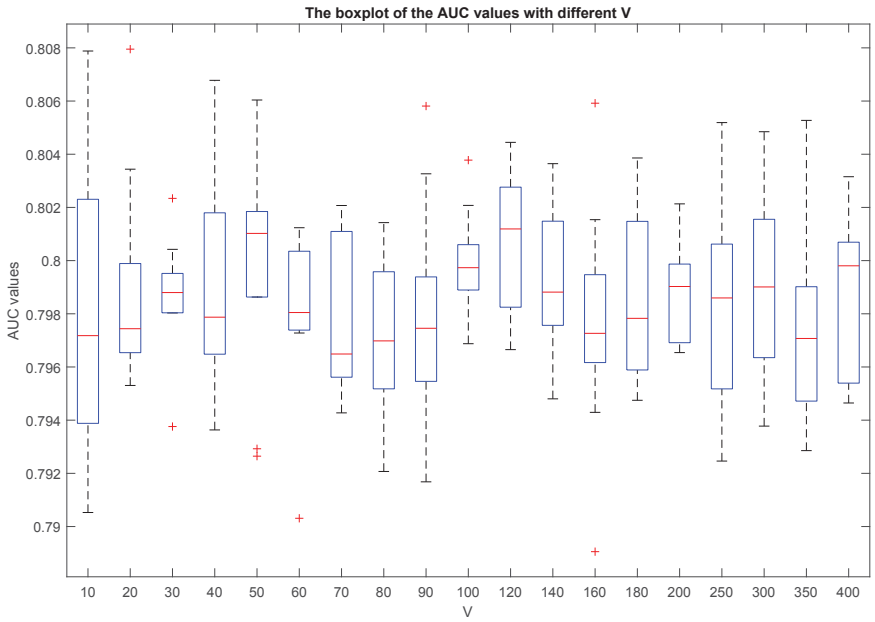


Figure 5.5: **The boxplot graph of the AUC values for the 5-fold cross validation experiments.** The x-axis is the value of $V$, and the y-axis is the corresponding AUC values. The changes of the AUC values with different $V$ are tiny. For a given $V$, the prediction results are stable.

### 5.3.4 Performance comparison and case studies

In comparison with two state-of-the-art disease-lncRNA association prediction methods LRLSLDA (Chen & Yan 2013) and LRLSLDA_ILNCSIM (Huang et al. 2016). Our leave-one-out cross-validation AUC performance is much better on the three datasets (Figure 5.6.). We noted that the source codes of these two existing methods are not available, but we implemented their algorithms for a fair comparison. Their datasets are not available either.

Our method with type 7 feature vector has a superior performance (AUC=0.8016, 0.8335 and 0.7527 on the three datasets) over the other three methods: the type 1 vector method (AUC=0.7889, 0.8266 and 0.7216), the LRLSLDA (AUC=0.6882, 0.7308 and 0.6346) and the LRLSLDA_ILNCSIM

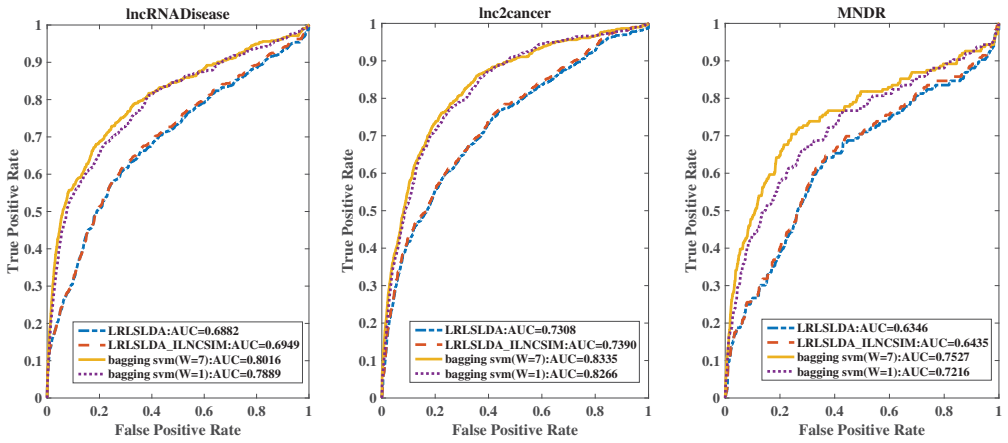


Figure 5.6: **The leave-one-out cross validation results based on three datasets with different methods.** Four methods were compared, our method with type 7 ($W$=7) feature and type 1 ($W$=1) feature, LRLSLDA method and the LRLSLDA_ILNCSIM method. Our type 7 method works best for all three datasets.

(AUC=0.6949, 0.7390 and 0.6435). We note that our type 1 vector needs just the accessible information such as disease genes and lncRNA sequences, but it can achieve close performance as the type 7 vector method did.

We also did the leave-one-disease-out cross-validation when assuming that all the related lncRNAs of a given disease are unknown. Then we computed the possibilities of the lncRNAs to be associated with the disease. The AUC value was used to test how are those already know related lncRNAs ranked comparing with the unknown ones. There are more than 40% (68 out of 162 diseases) of the diseases can achieve an AUC value higher than 0.9. The average AUC of all the diseases is 0.7356. This suggests that our method is capable of predicting disease-lncRNA associations even without knowing any association with a given disease.

We did an experiment to predict disease related lncRNAs using the known 454 positive samples and the 29840 unlabeled samples by PU learning. The predicted results were validated using two other datasets (166 lnc2cancer

samples and 29 MNDR samples overlap with the 29840 unlabeled samples). The ranking scores of the 29840 unlabeled samples and a ROC curve are plotted in Figure 5.7.

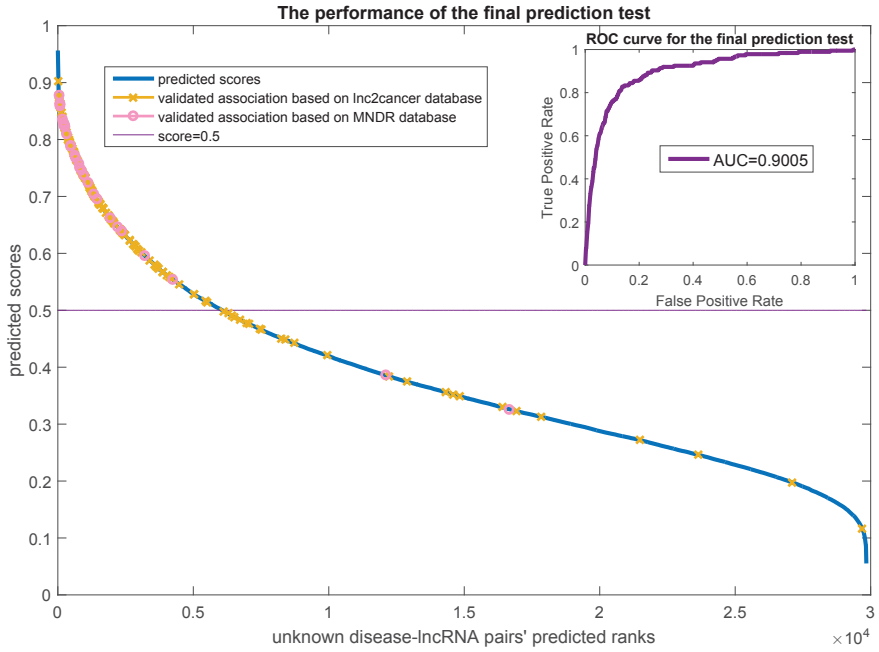


Figure 5.7: **The final prediction test on the lncRNADisease dataset.** The x-axis is the unknown disease-lncRNA pairs' predicted ranks. The y-axis are the predicted scores which means the possibilities of the samples to be positive. The predicted results were validated via the lnc2cancer and MNDR datasets. 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.

Figure 5.7 shows that most of the validated samples are ranked at good positions. We regarded those 166 lnc2cancer samples and 29 MNDR samples as positive while remaining unknown samples as negative and draw the ROC curve at the top right of Figure 5.7. It achieves the AUC value of 0.9005,

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

| Cancer type | Related lncRNA | Scores | Validated status |
|---|---|---|---|
| breast cancer | UCA1 | 0.8685 | validated by lnc2cancer |
| breast cancer | DLEU2 | 0.8375 | validated by literature |
| breast cancer | EPB41L4A-AS1 | 0.8356 | not validated |
| breast cancer | LINC00271 | 0.8297 | validated by literature |
| breast cancer | 7SK | 0.828 | validated by literature |
| prostate cancer | UCA1 | 0.922 | validated by lnc2cancer |
| prostate cancer | BCYRN1 | 0.8983 | not validated |
| prostate cancer | HOTAIR | 0.8952 | validated by lnc2cancer |
| prostate cancer | ZFAS1 | 0.881 | validated by literature |
| prostate cancer | BOK-AS1 | 0.88 | not validated |

which reveals that our prediction can always rank the positive samples well. We also did case studies for breast cancer and prostate cancer. Breast cancer is the leading type of cancer in women, accounting for 25% of all women cancer patients (McGuire 2016). Prostate cancer is the second most common type of cancer and the fifth leading cause of cancer-related death in men (McGuire 2016). We list in Table 5.2 top 5 lncRNAs that are (possibly) related to these two cancer types.

The most-top ranked lncRNA that is related to breast cancer is UCA1. This annotation has been already recorded in the lnc2cancer database. The second highest ranked lncRNA is DLEU2. In fact, DLEU2 is frequently deleted in malignancy (Lerner, Harada, Lovén, Castro, Davis, Oscier, Henriksson, Sangfelt, Grandér & Corcoran 2009). It functions as a critical host gene of the cell cycle inhibitory microRNAs miR-15a and miR-16-1. Both of these two microRNAs are related to breast cancer (Cittelly, Das, Salvo, Fonseca, Burow & Jones 2010). The 4th and 5th top-ranked lncRNAs LINC0271 (Delgado, Brandao & Narayanan 2014) and 7SK (Ji, Lu, Zhou & Luo 2014) are related to breast cancers. As to prostate cancer, two top-ranked lncRNAs UCA1 and HOTAIR have been actually stored in the

lnc2cancer database. In addition, the lncRNA ZFAS1 was recently reported to associate with the prostate cancer (Chen, Yang, Xie & Cheung 2018). These case studies support that our disease vector representation and PU learning methods are effective to prioritize disease related lncRNA genes.

## 5.4 Conclusion

In this chapter, we propose a novel disease vectorization method and apply it for the positive-unlabeled learning to predict and prioritize disease related lncRNA genes. It addressed part of my research question **Q1**. As was concluded in **Section 1.3 C3**, this part work has several contributions. Firstly, we observed the disease genes' chromosome distribution preference. Then, a disease is newly characterized by using the distribution properties of disease genes on the chromosome substructures and its related KEGG pathways to all the pathways. Our vectorization model can be applied to compute the disease similarities effectively. In addition, we proposed the bagging SVM based positive-unlabeled learning method for the classification of disease-lncRNA pairs. Testing on the benchmark datasets, our method can work better than the state-of-the-art methods. Especially, it can also work with only lncRNA sequences information or without known related association.

Future work has been planned to improve the performance of our vectorization model. First, more accurate disease genes will be collected as our model critically relies on the reliability of disease genes. Secondly, more information will be introduced to decrease the disease gene dependency such as the disease symptom, the disease semantics and so on. Furthermore, the relationship between the disease genes and lncRNA targets will be considered to extract more effective features to predict disease-lncRNA gene associations.

# Chapter 6

# CRISPR/Cas9 Cleavage Efficiency Regression Through Boosting Algorithms and Markov Sequence Profiling

## 6.1 Introduction

We have discussed in Chapter 1 that the key to good design of sgRNAs is to determine the spacer sequence by selecting a protospacer sequence complementary with the spacer's target sequence such that the cleavage (cleaving) efficiency is high. There are two critical prediction problems in the selection of sgRNAs. The first problem is the prediction of whether the sgRNA on-target cleaving efficiency is high or not. The subsequent problem is whether the sgRNA's off-target effect is low (Fu, Foden, Khayter, Maeder, Reyon, Joung & Sander 2013, Shen, Zhang, Zhang, Zhou, Wang, Chen, Wang, Hodgkins, Iyer, Huang et al. 2014, Kleinstiver, Pattanayak, Prew, Tsai, Nguyen, Zheng & Joung 2016). The first question is fundamental. This part work focuses on machine learning algorithms for assessing the cleaving efficiencies of candidate sgRNAs. The algorithms make regressions

on the numerical values of their cleavage efficiencies. The algorithms can be also turned to make binary classifications between high-efficiency and low-efficiency sgRNAs. The second question about the sgRNA off-target effects is closely linked to the first one. This question involves genome-wide number of genes which seems more complex, thus it is investigated in Chapter 7.

As was reviewed in Chapter 2, prediction algorithms have been recently proposed to identify efficient sgRNAs through characterizing their spacer sequence preferences (Doench et al. 2014, Xu et al. 2015, Wong et al. 2015, Kaur et al. 2016, Moreno-Mateos et al. 2015), thermodynamics features (Doench et al. 2014, Wong et al. 2015) and structure features (Wong et al. 2015). The sequence features are widely adopted because many nucleotide preference phenomena have been observed. For example, nucleotides distal to the PAM were found to be dominated by the guanine enrichment, while the remaining nucleotides are characterized by the cytosine enrichment (Moreno-Mateos et al. 2015). These nucleotide preference properties have been exploited to differentiate efficient sgRNAs from those inefficient ones by machine learning methods such as support vector machine (SVM) (Doench et al. 2014, Wong et al. 2015, Kaur et al. 2016, Rahman & Rahman 2017). In particular, a regression method (Doench et al. 2016) has been proposed to predict the numerical values of the cleaving efficiencies for candidate sgRNAs. Its novel idea is a Rule Set 2 (RS2) for predicting the on-target activities of sgRNAs. Different from the previous classification methods, this regression model also uses some new features such as cutting position features and the two nucleotides in the N and N positions relative to the PAM 'NGGN'. Though RS2 achieved remarkable performance, there still exists large space for improving the performance.

This work introduces a two-step averaging method (TSAM) for the prediction of sgRNA cleaving efficiencies. At the first step, a boosting regression model is trained on the conventional feature space of sgRNAs to map these sgRNAs to their cleaving efficiency scores. At the second step, we use Markov sequence profiles of sgRNAs as new features together with

important features selected by the boosting algorithm to train a non-linear SVM to make regression again on the cleaving efficiencies. The two scores are then averaged as the predicted cleaving efficiencies of these sgRNAs.

Both the boosting algorithm and the Markov sequence profiling have the same aim to exploit important characteristic features of sgRNAs to improve the prediction performance but at different aspects. Literature methods already proposed a large number of features to describe sgRNAs. However, not all of them are effective for the prediction of the cleavage efficiencies. The newly introduced Markov sequence features can capture the global sequence characteristics of sgRNAs which are different from the conventional position-specific preferences (Doench et al. 2014, Wong et al. 2015, Kaur et al. 2016, Doench et al. 2016). The boosting algorithm, XGBoost (Chen & Guestrin 2016), is a scalable end-to-end tree boosting system that can rank the feature importance during the training process. XGBoost is also a state-of-the-art regression algorithm with better performance than the traditional gradient boosting trees (Doench et al. 2016), having a wider range of applications (Zhang, Ai, Chen, Yin, Hu, Zhu, Zhao, Zhao & Liu 2017, Torlay, Perrone-Bertolotti, Thomas & Baciu 2017). Furthermore, our two-step averaging strategy underlines a complementary nature of the boosting regression approach and the SVM regression approach. From our experiments, the regression results of XGBoost and SVM are always different. It is good to integrate the two regression results to improve the prediction performance on the sgRNA cleaving efficiencies.

Markov sequence profiles of a sgRNA are extracted through a profile Hidden Markov Model (pHMM). It works by converting a multiply sequence alignment for sequences from a known family into a position-specific scoring system (Eddy 1998). This system can be used to evaluate whether a new sequence is a homologous sequence of this sequence family. This method has been leveraged to address many other biological sequence related bioinformatics problems (Karplus, Barrett & Hughey 1998, Schliep, Schönhuth & Steinhoff 2003, Wheeler, Clements, Eddy, Hubley, Jones, Jurka,

Smit & Finn 2013, Huo, Zhang, Huo, Yang, Li & Yin 2017). In this work, sgRNA sequences are first grouped into sub-families in accordance with their efficiency scores. Then, probabilities of a given sgRNA being a homologous sequence for each sub-families are formed as a multi-pHMM vector for characterizing the global features of sgRNA sequences. An SVM regressor trained with only pHMM properties can obtain similar mean Spearman correlations comparing with the state-of-the-art methods. Hence, we decided to combine pHMM features with the top-ranked features of XGBoost to train the second-step SVM regressor for a better performance.

The performance of our TSAM is compared with the state-of-the-art regression methods such as RS2 (Doench et al. 2016) and CRISPRscan (Moreno-Mateos et al. 2015). On Doench's FC dataset (human and mouse sgRNAs), TSAM obtained a mean Spearman correlation of 0.583, better than RS2's 0.522. On the RES dataset (human sgRNAs) and the FC+RES dataset, TSAM achieved mean Spearman correlations 0.530 and 0.567 respectively, better than RS2's 0.455 and 0.510. On the dataset which was used by CRISPRscan containing 1020 zebrafish sgRNA sequences, TSAM can achieve a competitive Pearson correlation of 0.49 (comparing with CRISPRscan's 0.45). Our two-step regression approach was converted into a binary classification method to distinguish between high-efficiency and low-efficiency sgRNAs. The classification performance on the benchmark datasets also outperforms the state-of-the-art methods. For instance, the mean AUC of the three-fold cross validation on Xu's ribosomal dataset (Xu et al. 2015) is 0.896, much exceeding Xu's 0.843. For the cross-gene validation and cross-platform validation, our performances are 0.813 and 0.840 respectively, better than Xu's 0.778 and 0.757.

Haeussler et al. (Haeussler et al. 2016) advised that the performance of an on-target efficiency prediction model is strongly dependent on whether the guide RNA is expressed from a U6 promoter or it is transcribed in vitro with the T7 promoter. To compare the performance of TSAM with the state-of-the-art methods on datasets of different expression systems, we

collected abundant datasets from (Haeussler et al. 2016) to test two specified versions of TSAM. One is named TSAM_U6, which was trained with the FC+RES dataset as input, in which the guide RNAs are all transcribed from the U6 promoter. The dataset CRISPRScan containing guides expressed from the T7 promoter in vitro was used to build the second predictor named TSAM_T7. The results confirmed that our TSAM can always achieve better performances on both of the U6 and T7 promoter datasets.

Our case studies are related to the optimal sgRNAs selected for gene therapies to cure the retinitis pigmentosa and X-linked chronic granulomatous disease (Yu, Mookherjee, Chaitankar, Hiriyanna, Kim, Brooks, Ataeijannati, Sun, Dong, Li et al. 2017, De Ravin, Li, Wu, Choi, Allen, Koontz, Lee, Theobald-Whiting, Chu, Garofalo et al. 2017). The highly efficient sgRNAs recommended by our method can well match with those sgRNAs which had been validated by wet lab experiments and domain experts. This partly proves the effectiveness of our prediction tool, and illustrates the great potential of our method for practical use. The web-server can be freely visited from the site: http://www.aai-bioinfo.com/CRISPR/. The off-line tool can be downloaded from the website: https://github.com/penn-hui/TSAM.

## 6.2 Materials and Methods

### 6.2.1 High throughput genome engineering datasets for building the regression and classification models

We tested the algorithms on total 11 datasets. Three datasets from (Doench et al. 2016) were downloaded to build our TSAM regression model. The three datasets are named: the FC dataset which contains 1841 sgRNAs with the flow cytometry (FC) method detecting the knockdowns; the RES dataset which contains 2549 sgRNAs with their knockdown efficiencies measured through drug resistance detection; and the combined dataset

(FC+RES). We removed 10 sgRNAs from the FC dataset because of their ambiguous mapping to the reference genome (Fusi, Smith, Doench & Listgarten 2015). Doench's paper reported that there are 1831 curated sgRNAs in the FC dataset, however, there are only 1830 unique sgRNAs from their supplementary materials. Furthermore, 1020 sgRNAs for cleaving zebrafish genome sequences were acquired from (Moreno-Mateos et al. 2015). Different from FC and RES, where the guides are transcribed from U6 promoters in cells, this zebrafish dataset contains the guides expressed from T7 promoters in vitro. As the cutting efficiency measurement methods are distinct, separate models are trained and evaluated on these different datasets. More details of these four datasets are listed at the first 4 rows of **Table 6.1**.

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

| Name | validation type | sample size | literature |
|---|---|---|---|
| FC | logo[1] | 1830 | (Doench et al. 2016) |
| RES | logo | 2549 | (Doench et al. 2016) |
| FC+RES | logo | 4379 | (Doench et al. 2016) |
| CRISPRScan | ShuffleSplit | 1020 | (Moreno-Mateos et al. 2015) |
| Xu_ribo | threefold | 731H,438L[2] | (Xu et al. 2015) |
| Xu_non-ribo | inter-geneset[3] | 671H,237L | (Xu et al. 2015) |
| Xu_mouse | inter-platform[4] | 830H,234L | (Xu et al. 2015) |
| Xu_inde1 | independent[5] | 52H,25L | (Xu et al. 2015) |
| Xu_inde2 | independent | 110H,110L | (Xu et al. 2015) |
| Chari_spCas9 | tenfold | 133H,146L | (Chari et al. 2015) |
| Chari_stlCas9 | tenfold | 82H,69L | (Chari et al. 2015) |

[1] regression, leave-one-gene-out cross-validation

[2] classification, where H for efficient and L for inefficient

[3] trained on Xu_ribo and tested on Xu_non-ribo

[4] trained on Xu_ribo + Xu_non-ribo and tested on Xu_mouse

[5] trained on Xu_ribo + Xu_non-ribo + Xu_mouse and tested on Xu_inde1

In the test of whether our TSAM can address the problem of classifying

sgRNAs into high-efficiency or low-efficiency ones, five datasets from (Xu et al. 2015) were downloaded including three datasets for three-fold cross-validation, inter-geneset validation and inter-platform validation, and two independent test sets (directly from the authors) for evaluation and comparing the performances of different methods. The details are listed at the 5th to 9th rows of **Table 6.1**.

To compare with Chari's sgRNA Scorer (Chari et al. 2015), their datasets were obtained from the supplementary files of the published paper (shown at the last two rows of **Table 6.1**). Chari et al. tested their method on two datasets: a 133 high-activity vs 146 low-activity sgRNA dataset for the assessment of spCas9 system, and an 82 high vs 69 low sgRNA dataset for the stlCas9 system (from *Streptococcus thermophilus*, where its PAM is NNAGAAW). All the adopted datasets can be found from our **Supplementary file 16**.

### 6.2.2 Features for building the regression and classification models

#### Conventional sequence features

Here, an sgRNA sequence is always referred to as the protospacer sequence corresponding to the spacer and its upstream to the PAM. To extract some similar features as used by RS2 (Doench et al. 2016), we similarly extended the sequences to 30nt in length, namely $N_4N_{20}NGGN_3$ (N represents any nucleotide, the first 4nt and the last 3nt are also extracted together with the original 20nt spacer and the PAM NGG). An sgRNA sequence is denoted as $S = s_1s_2s...s_i...s_{30}$, where $s_i \in \{A, G, C, T\}$.

**Nucleotide composition features:** The number of each single nucleotide (e.g., how many 'A') in $S$ is counted, and each characterized as an order 1 nucleotide composition (nc1) feature. Similarly, the number of each dinucleotide or trinucleotide (e.g., how many 'AA' or 'AAA' in $S$) is counted, and each characterized as an order 2 or order 3 nucleotide composition feature

(nc2, or nc3). The counts of the dinucleotides and the trinucleotides were computed by a sliding window mechanism.

**Position specific nucleotide binary features:** An order 1 position specific nucleotide binary feature (psnb1), at a given position, is initialized as a vector (0, 0, 0, 0). The first element represents whether the nucleotide at this position is 'A'. If yes, change the 0 to be 1. The second element represents the status of 'G', the third for 'C' and the forth for 'T'. For example, if at position 1, the nucleotide is 'A' then, this vector is (1, 0, 0, 0), or if the nucleotide is 'C', this vector is (0, 0, 1, 0). Similarly, an order 2 position specific nucleotide binary feature (psnb2) and order 3 position specific nucleotide binary feature (psnb3) are established in the same way, where every dinucleotide and trinucleotide are used as an element of the 16-dimensional vector and 64-dimensional vector at a given position.

**GC features:** Each of these features describes the counts of how many 'G' or 'C' in $S$ (named GC counts features), or the percentage of 'G'+'C' in $S$ (named the GC percent feature).

### Thermodynamic features

The melting temperatures of sgRNA sequences at different regions were computed with the Biopython Tm_staluc function (Cock, Antao, Chang, Chapman, Cox, Dalke, Friedberg, Hamelryck, Kauff, Wilczynski et al. 2009, Le Novere 2001). We considered the following regions as features: the whole 20nt spacer (TMr1), the core region (12nt adjacent to PAM, TMr2), the non-core region (the remaining 8nt of the 20nt spacer, TMr3), the whole 30nt extended sgRNA sequence (TMr4), the 5nt adjacent to PAM (TMr5), the 8nt proximal to the previous 5nt (TMr6) and another 5nt next to the middle 8nt (TMr7). The last four regions have been used by RS2 (Doench et al. 2016).

### Cutting position related features

Cutting positions relative to protein sequences have been used to improve the performance on the prediction of sgRNA cleaving efficiencies (Doench et al. 2016). In this work, we considered the cutting position to the genome sequence (cut_geno), to the transcript sequence (cut_trans) and to the protein sequence (cut_pro) as three features. Meanwhile, the percentage of the cutting length was considered as a feature computed as the length from the start of the sequence to the cut position divided by the whole sequence length (denoted as cut_per_geno, cut_per_trans, and cut_per_pro respectively). The gene's genome sequence, transcript sequence, protein sequence and the detail exon, intron, 5'UTR and 3' UTR sequences were downloaded from the ensembl database (Hubbard, Barker, Birney, Cameron, Chen, Clark, Cox, Cuff, Curwen, Down et al. 2002) for the mapping of these cutting positions. The gene's start coordination was normalized to be 1 for calculating feature values of cut_geno, cut_trans, cut_per_geno and cut_per_trans. Features cut_trans, cut_pro, cut_per_pro and cut_per_trans were set to be a value of 0 if the sgRNA cut in an intron region. Features cut_pro and cut_per_pro were also set to be a value of 0 if the sgRNA cut at non-coding regions.

### Profile hidden Markov model (pHMM) features of sgRNA sequences

It is the sgRNA sequence as a whole that can truly determine its cutting efficiency. Here, the global features of an sgRNA sequence are extracted through a profile hidden Markov model (Eddy 1998). We hypothesized that those sgRNAs with similar cutting efficiencies should contain more sequence similarities and vice versa. Thus, these sgRNAs can be grouped into subfamilies where the efficiencies of the sgRNAs in each group are similar. Then, if a new sequence belongs to a subfamily, its cutting efficiency may also similar to its homologous sequences. The pHMM was adopted to solve this homologous sequence searching problem, where the pHMM properties were used to characterize the sgRNA sequences.

A pHMM is usually used for modeling multiple sequence alignments and it can provide a probabilistic model for comparing new sequences to the multiple alignments (Durbin, Eddy, Krogh & Mitchison 1998). Traditional pHMM can be described with an HMM composed of a state set $S = \{Begin, Match, Insert, Delete, End\}$ and an alphabet of symbols $\mho = \{e_1, e_2, \dots\}$ that are emitted by the non-silent states (usually are Match and Insert states). After training on a sequence family (a protein family or a set of homologous gene sequences), a transition probability matrix and an emission probability matrix can be constructed to depict the transitions between the states and the emission status of the non-silent states. For a given sequence, a log-sum-of-odds score describing the probability of the pHMM generating it can be computed by the *Viterbi* algorithm (Forney 1973). Please be referred to (Eddy 1998, Durbin et al. 1998) for more details about pHMM.

Most of the high throughput experiments fixed the spacer length as 20nt. Thus, the spacer sequences here were set to be well aligned with the fixed length 20 (there is no Insert or Delete state but only Match status), where the pHMM is a so-called BLOCK-style ungapped motif (Eddy 1998). Two sets of symbols were permitted to be emitted at the Match state, i.e., a single nucleotide set $\mho_1 = \{A, G, C, T\}$ and a dinucleotide set $\mho_2 = \{AA, AG, AC, AT, \dots, TA, TG, TC, TT\}$. To avoid the emission probability of zero, we add pseudocounts to the observed counts. Therefore, the emission probability $e_i$ is calculated as $e_M(e_i) = \frac{count(e_i)+pu}{count(all)+pd}$, where $pu$ and $pd$ are the pseudocounts for the observed count of each emitted symbol and all the emissions.

Suppose there is a set of sgRNAs $Sg = \{sg_1, sg_2, \dots, sg_j, \dots, sg_m\}$ with known efficiencies $Ef = \{ef_1, ef_2, \dots, ef_j, \dots, ef_m\}$, $ef_j \in [0, 1]$. For an sgRNA $\ell$, its pHMM properties are extracted by the following two steps:

- **Step1: Grouping $Sg$ into $k$ sub-families and training their pHMMs.** Separating $Sg$ into $k$ sub-families $Sf = \{sf_1, sf_2, \dots, sf_x, \dots, sf_k\}$, where each of them has an efficiency range, e.g., $ef(sf_x) \in$

$[0.1, 0.2)$. For $sf_x \in Sf$ and a given emission symbol type $t$, a pHMM can be trained with its sequences. These pHMMs are denoted as $H^t = \{h_1^t, h_2^t, \ldots, h_x^t, \ldots, h_k^t\}$.

- **Step2: Extracting $\ell$'s pHMM vector.** For sgRNA $\ell$, the probability $hf_x^t$ generated by $h_x^t$ is computed by the *Viterbi* algorithm, and $\ell$ is characterized by a vector $Hf_\ell^t =< hf_1^t, hf_2^t, \ldots, hf_x^t, \ldots, hf_k^t >$, where $t = \mho_1, \mho_2$.

Here both of the two emission symbol sets $\mho_1$ and $\mho_2$ are used, which can produce two vectors for sgRNA $\ell$, i.e., $Hf_\ell^{\mho_1}$ (pHMMe1) and $Hf_\ell^{\mho_2}$ (pHMMe2).

### 6.2.3 Procedures for training our TSAM

Our TSAM cleaving efficiency regression model is built by four main steps. Firstly, all the features are created. Then, an XGBoost regressor is trained with some selected primary features to estimate the first-step scores. The features' importances are evaluated simultaneously. Later, the most important features are combined with the pHMM features to optimize an RBF SVM regressor. Then the second-step scores are calculated. At last, the first-step scores and the second-step scores are averaged as the final scores for the regression. **Figure 6.1** shows the flowchart to construct TSAM.

To get the best training performance on the dataset FC, the XGBoost and SVM regression methods were both optimized by the leave-one-gene-out cross-validation for the best parameters. The best parameters were fixed when these two regression methods were used to generate leave-one-gene-out cross-validation performance on the RES dataset or on the FC+RES dataset. To have a fair performance comparison with (Moreno-Mateos et al. 2015) on the CRISPRScan dataset, our regression methods were also optimized by the same Shuffle-Split cross-validation as (Moreno-Mateos et al. 2015) did.

We also note that there is a pre-evaluation process to select important features from the initial feature set for optimizing the XGBoost regressor.

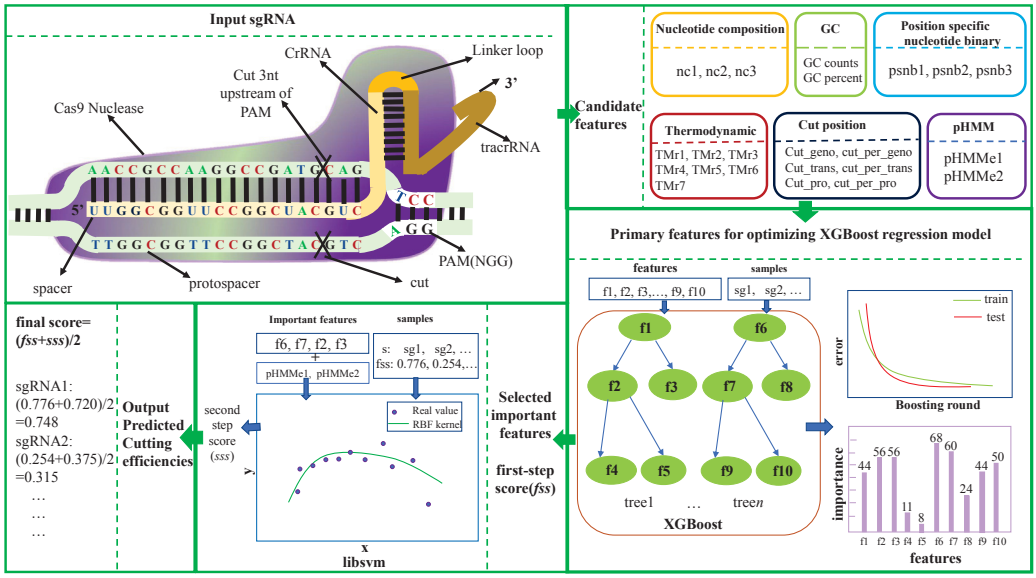


Figure 6.1: **The flowchart to construct TSAM for predicting sgRNA cleavage efficiencies.** This flowchart contains four main steps: at first 6 types of initial features are created; in the second step, primary features are selected from the initial feature set to optimize an XGBoost regressor and output 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 ($(fss + sss)/2$).

This process is implemented by the backward elimination strategy (Mao 2004) with default parameters for XGBoost. During each fold of the cross-validation, the selected features are assigned with feature importance to weight their contributions for optimizing the regressor.

The features that work well for SVM (e.g., the pHMMe1 and pHMMe2 according to our results) are combined with the boosting selected top-$K$ important features to train an RBF kernel SVM regressor (libSVM v3.22 (Chang & Lin 2011)). As the features' importance are evaluated during each cross-validation fold, the final selected important features are the union of

the top-$K$ ones from all the folds. This SVM regressor predicts the second-step scores for the sgRNAs. The details of determining the parameters for regressors and features are described in the **Supplementary file 15** (see Appendix C).

## 6.3 Results

We first report the cleavage preferences of sgRNAs as revealed by XGBoost and explain how these preferences are different from literature observations. Then, we report excellent regression performance achieved by integrating XGBoost and SVM. These results and analysis are mainly focused on the dataset FC. After that, we present comparison results between our method and the state-of-the-art methods to demonstrate the superior performance on the sgRNA cleavage efficiency regression by our method. At last, two case studies are presented to illustrate the effectiveness of our method for practical use in gene therapies.

### 6.3.1 Nucleotide and cleavage preferences of highly efficient sgRNAs as revealed by the boosting algorithm

Some interesting nucleotide preferences of the highly efficient sgRNAs are revealed by the XGBoost algorithm on the FC dataset (see **Figure 6.2)**. A highly efficient sgRNA is always a sequence of relatively lower melting temperature at the middle of the spacer, in comparison with those of low efficiencies (a mean value 8.84 for the highly efficient sgRNAs that are ranked at the top 20% of the 1830 sgRNAs according to their actual efficiencies vs 13.11 for the low efficient sgRNAs ranked at the bottom 20%, p-value=1.04E-09 under the two-sample Kolmogorov-Smirnov test (Lilliefors 1967)). Also, the highly efficient sgRNAs prefer to cut at the 5'-end closer part of a gene (a mean value of cut_per_geno is 41.56% for the highly efficient sgRNAs vs

46.61% for the low efficient sgRNAs, p-value=1.51E-04). In addition, the nucleotide composition of the highly efficient sgRNAs and the low efficient sgRNAs exhibits a distinct divergence: the highly efficient sgRNAs have more 'A' (on average 6 for the highly efficient sgRNAs vs. 5 for the low efficient ones, p-value=2.83E-09), but less 'G' (on average 10 for the highly efficient vs. 11 for the low efficient, p-value=5.06E-03), 'GG' (on average 4 for the highly efficient vs. 5 for the low efficient, p-value=3.23E-08) and 'GGG' (on average 1 for the highly efficient vs. 2 for the low efficient, p-value=6.51E-07).

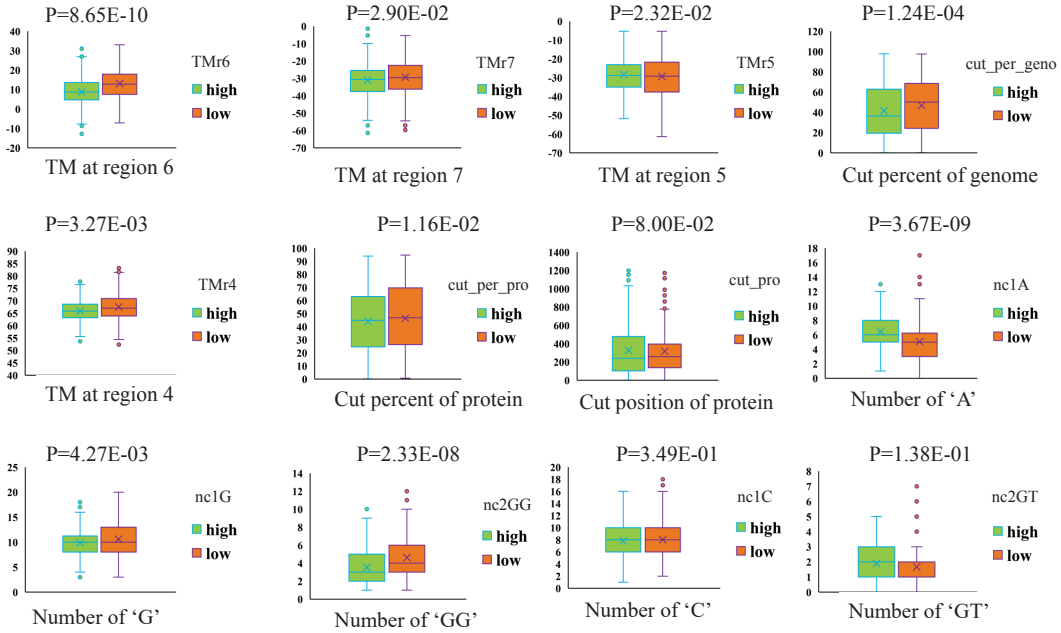


Figure 6.2: **Top 12 important features and analysis on the nucleotide and cleavage preferences.** Y-axis shows the feature values. The feature names are placed under the x-axis and their symbols are placed at the top right panel of the subplots. These features are ranked by their importance. Type "high" 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.

(Doench et al. 2016) have reported that the three types of features that contribute substantially to the efficiency prediction are: position-independent

counts of single and dinucleotides, location of the sgRNA within the protein, and melting temperatures at different regions (having Gini importance of 16%, 13% and 11% respectively). By our boosting algorithm, these three types of features constitute the top 25 sub-features whose importance are higher than 100. Different conclusions are drawn as follows. First, the melting temperatures at different regions are the best features (with a mean importance 542.64), then the cutting position related features are (with a mean importance 424.41), followed by the nucleotide composition related features (with an average importance 136.59). Meanwhile, the cutting percent relative to genome DNA sequence (cut_per_geno, not applied by RS2) is better than the cutting percent relative to protein (cut_per_pro) and the cutting position at the protein (cut_pro) (importance are 503.89, 399.44 and 369.89 respectively). The divergences of the values for cut_per_pro and cut_pro between the high and low efficient sgRNAs are not as significant as that of the cut_per_geno (p-value=1.39E-02, 8.41E-02 and 1.51E-04 respectively).

The regression performance on the cleaving efficiencies by our XGBoost is better than Doench et al.'s RS2. We obtained a mean Spearman correlation 0.562, but RS2 obtained only 0.522 on the FC dataset. This is why conclusions on the nucleotide preferences of highly efficient sgRNAs are different between these two methods. We note that our XGBoost regressor did not use all the features but only important features such as TMr4-TMr7, nc1, nc2, nc3, psnb1, psnb2, GC counts, GC percent, cut_per_geno, cut_pro and cut_per_pro (form 677 dimensions in total). More details about the XGBoost regression parameter settings and the features can be found at our **Supplementary file 15**.

### 6.3.2 Further performance improvement by integrating pHMM properties

The pHMM properties (combining the pHMMe1 and pHMMe2) can be used to build an SVM regressor to achieve fairly good performance, where a

mean Spearman correlation 0.519 was obtained. Adding the top-ranked important features evaluated by the former boosting can further improve the SVM regressor's mean Spearman correlation to 0.559 which is superior to Doench's methods (RS2's mean Spearman correlation=0.522 and L1-Regression's mean Spearman correlation=0.513). If the pHMM properties were removed from this strong SVM regressor, the performance dropped about 0.01. This implies that the pHMM properties are indispensable to construct our excellent SVM regressor.

The proposed TSAM obtained a mean Spearman correlation 0.583 which is much better than Doench's methods. It also improves the mean Spearman correlation of our XGBoost regressor by 0.021, benefited from its integration with the SVM regressor trained on the pHMM properties and other significant features. The SVM regressor alone also achieved better performance than Doench's methods but worked not as well as TSAM. This proves that the XGBoost regressor and the SVM regressor can predict the sgRNA's cutting efficiencies cooperatively. The parameter optimization process is described in **Supplementary file 15**.

### 6.3.3 Results on 11 benchmark datasets comparing with the state-of-the-art methods

Four benchmark datasets were used to evaluate the performance of our proposed TSAM. The performance was compared with the following state-of-the-art methods: Doench et al's RS2, L1-Regression methods (implemented by this work) (Doench et al. 2016), and the CRISPRscan method (Moreno-Mateos et al. 2015). Our TSAM improves the mean Spearman correlation by more than 0.05 comparing with RS2 and L1-Regression on the FC, RES and the FC+RES datasets (under the leave-one-gene-out evaluation framework), and improves the mean Pearson correlation by about 0.04 comparing with CRISPRscan (under the same Shuffle-Split evaluation framework) on the sgRNAs dataset for cutting zebrafish genome sequences. The detailed results are presented at the first four rows of **Table 6.2**.

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

| Methods | regression performance on | | | |
|---|---|---|---|---|
| | FC | RES | FC+RES | CRISPRscan |
| RS2 | 0.522 | 0.455 | 0.510 | - |
| L1-Regression | 0.513 | 0.468 | 0.505 | - |
| CRISPRscan | - | - | - | 0.45 |
| **TSAM** | **0.583** | **0.530** | **0.567** | **0.488** |
| TSAM-MT1 | 0.565 | 0.441 | 0.531 | 0.475 |
| TSAM-MT2 | 0.575 | 0.493 | 0.555 | 0.477 |

In the further evaluation of TSAM, we have conducted the cross-dataset test. We trained TSAM on the FC dataset, and then the sgRNAs belonging to the 8 genes in the RES dataset were adopted as 8 independent test sets. The mean Spearman correlation by our regression is 0.431, which is much better than the performance by Doench's methods (0.397 by RS2 and 0.383 by the L1 regression). On the 8 genes, we obtained higher Spearman correlations on 6 of them than Doench's RS2 and L1 regression methods. When TSAM was trained with the 2549 sgRNAs from the RES dataset and tested on the 9 genes from the FC dataset, the mean Spearman correlation was 0.551 for TSAM, while Doench's RS2 and L1 regression obtained only 0.508 and 0.493 respectively. As expected, we obtained better Spearman correlations than Doench's methods on 7 of the 9 genes.

We have conducted a stricter performance evaluation for TSAM to satisfy practical use conditions especially assuming the cutting position features are not accessible. For this performance test, we modified TSAM as two Mutation Types (MT): TSAM-MT1 and TSAM-MT2. TSAM-MT1 was trained without cutting position features (674-d, deleting the cut_per_geno, cut_pro and cut_per_pro), and TSAM-MT2 was trained without the cutting position related to the protein features (675-d, without cut_pro and cut_per_pro). The performances of these two variant methods are shown

in the last two rows of **Table 6.2**. It is understood that the cutting position features can significantly affect the performance of our TSAM on the RES dataset. Except for one case testing on the RES dataset, our methods obtained much better performance than the state-of-the-art methods.

Our TSAM regression method can be easily converted for a binary classification approach to the distinction between highly efficient sgRNAs and low efficient ones. The steps are as follows. First, XGBoost was optimized to output feature importance scores (classification with the binary logistic function). Then, the important features were combined with the pHMM properties to train an SVM classifier with an RBF kernel (the pHMM group is set as 2, such as positive sample group and the negative sample group, probabilities as output). Then the classifier was tested on 7 datasets including 5 datasets for cross-validation and 2 independent test sets. The other classifiers (Doench et al. 2014, Xu et al. 2015, Chari et al. 2015) were also optimized with the corresponding validation types in **Table 6.1**. The cross-validations were repeated 10 times and the performances were averaged as the final performance. Then the classification performances were weighted by Matthews correlation coefficient (MCC) (Matthews 1975), F1, AUC and Accuracy which are all shown in **Table 6.3**.

The variant method TSAM-MT1, instead of TSAM itself, was applied to test the performance on the Chari_stlCas9 dataset. The reason is that the PAM of the sgRNAs was defined as 'NNAGAAW' but not the 'NGG' motif. Thus the cutting position features could not be defined. We can see that TSAM-MT1 can outperform the state-of-the-art methods as well for the binary classification of sgRNAs. More comparison results are provided at **Supplementary file 15**.

### 6.3.4 Performance of TSAM on more datasets related to the U6 and T7 expression system

We used the datasets from (Haeussler et al. 2016) to confirm that the proposed TSAM can work better than RS2 when the guide RNAs are

Table 6.3: **Performance comparison between our method and the state-of-the-art methods for the binary classification of sgRNAs.**

| Method | dataset | MCC | F1 | AUC | Accuracy |
|---|---|---|---|---|---|
| TSAM | Xu_ribo | 0.640 | 0.871 | 0.896 | 0.834 |
| Xu et al.'s | Xu_ribo | - | - | 0.843 | - |
| TSAM | Xu_non-ribo | 0.505 | 0.884 | 0.813 | 0.822 |
| Xu et al.'s | Xu_non-ribo | - | - | 0.778 | - |
| TSAM | Xu_mouse | 0.508 | 0.891 | 0.840 | 0.830 |
| Xu et al.'s | Xu_mouse | - | - | 0.757 | - |
| TSAM | Xu_inde1 | 0.311 | 0.800 | 0.798 | 0.714 |
| Xu et al.'s | Xu_inde1 | - | - | 0.729 | - |
| Doench et al. | Xu_inde1 | - | - | 0.648 | - |
| TSAM | Xu_inde2 | 0.433 | 0.748 | 0.779 | 0.700 |
| Xu et al.'s | Xu_inde2 | - | - | 0.711 | - |
| Doench et al.'s | Xu_inde2 | - | - | 0.583 | - |
| TSAM | Chari_spCas9 | 0.551 | 0.758 | 0.859 | 0.772 |
| Chari et al.'s | Chari_spCas9 | - | - | - | 0.732 |
| TSAM-MT1 | Chari_stlCas9 | 0.718 | 0.865 | 0.930 | 0.855 |
| Chari et al.'s | Chari_stlCas9 | - | - | - | 0.815 |

expressed from U6 and better than CRISPRscan when the expression system is T7.

### Comparison on datasets from the U6 expression system

We compared the prediction performance of TSAM_U6 and RS2 on 7 big datasets containing sgRNAs for cutting human or mouse genomes. Both TSAM_U6 and RS2 are trained on the FC+RES dataset, where the sgRNAs are expressed from U6 promoters in cells. The Spearman correlation are shown in **Table 6.4**.

We can see that for all the seven datasets each containing more than 1000 sgRNAs, our TSAM_U6 achieved about 3% more the Spearman correlation than RS2.

Table 6.4: **Spearman correlation of TSAM, RS2 and CRISPRscan tested on datasets from U6 or T7 expression systems.**

| U6 expression system | | | | | |
|---|---|---|---|---|---|
| dataset | size | genome | literature | TSAM_U6 | RS2 |
| Wang/Xu HL60 | 2076 | *Human* | *Wang et al.* (2014) | **0.517**[1] | 0.485 |
| Chari 293T | 1234 | *Human* | *Chari et al.* (2015) | **0.382** | 0.381 |
| Hart Rpe | 4214 | *Human* | *Hart et al.* (2015) | **0.309** | 0.281 |
| Hart Hct116-2 Lib 1 | 4239 | *Human* | *Hart et al.* (2015) | **0.416** | 0.384 |
| HartHelalib1 | 4256 | *Human* | *Hart et al.* (2015) | **0.388** | 0.353 |
| HartHelalib2 | 3845 | *Human* | *Hart et al.* (2015) | **0.394** | 0.359 |
| XuKBM | 2076 | *Human* | *Xu et al.* (2015) | **0.540** | 0.512 |
| T7 expression system | | | | | |
| dataset | size | genome | literature | TSAM_T7 | CRISPRscan |
| Eschstruth Zebrafish | 17 | *Zebrafish* | *Haeussler et al.* (2016) | **0.224** | -0.0043 |
| Varshney Zebrafish | 102 | *Zebrafish* | *Varshney et al.* (2015) | **0.363** | 0.262 |
| Gagnon Zebrafish | 111 | *Zebrafish* | *Gagnon et al.* (2014) | **0.410** | 0.357 |
| Shkumatava Zebrafish | 162 | *Zebrafish* | *Haeussler et al.* (2016) | **0.292** | 0.258 |
| Teboul Mouse In Vivo | 30 | *Mouse* | *Haeussler et al.* (2016) | **0.565** | 0.426 |

[1] For each dataset, the highest Spearman correlation is in bold

### Comparison on datasets from T7 expression system

Another 5 datasets whose sgRNAs are expressed from T7 promoters were used to compare the performances between TSAM_T7 and CRISPRscan. Both of these two predictors were trained with the CRISPRScan dataset and the sgRNAs in this dataset are expressed from a T7 promoter in vitro. The Spearman correlations are listed in **Table 6.4**. Again, the proposed TSAM_T7 achieved 10% more the Spearman correlation on 3 out of 5 datasets and about 5% more on the remaining two datasets than the best existing predictor CRISPRscan for this type of expression system. See our **Supplementary file 15** and **Supplementary file 17** for detailed results and the applied datasets.

### 6.3.5 Case study: designing sgRNAs for gene therapy

CRISPR/Cas9 system is a very promising genome engineering tool for curing genetic diseases (Men, Duan, He, Yang, Yao & Wei 2017). In the understanding of whether TSAM can recommend reasonable sgRNAs for practical use, we conducted case studies for recommending sgRNAs to treat retinitis pigmentosa and X-linked chronic granulomatous disease. Gene editing investigations on these two diseases have been successfully undertaken by domain experts recently (Yu et al. 2017, De Ravin et al. 2017).

Yu et al. (Yu et al. 2017) attempted to knockdown gene Nrl to prevent retinal degeneration in a mouse model and suggested adopting CRISPR/Cas9-mediated NRL disruption in rods as a promising treatment option for patients with retinitis pigmentosa. For our prediction, the genome sequences of mouse Nrl gene were downloaded from Ensembl database under the transcript id ENSMUST00000062232.13. Total 138 potential spacer sequences were found with the PAM 'NGG'. Among these 138 candidate sgRNAs, the cleavage efficiencies of those sgRNAs cutting at the coding region were predicted by our TSAM method. If considering just the cutting efficiency, the 3 top-ranked sgRNAs' spacer sequences are 5'-ATGCCTGGCTCACTGAAGGT-3' (s1, cut efficiency=0.850), 5'-GTATGGTGTGGAGCCCAACG-3' (s2, cut efficiency=0.801) and 5'-CACA-GACATCGAGACCAGCG-3' (s3, cut efficiency=0.762). Yu's work proposed to use 5 candidate sgRNAs (denoted NT1 to NT5). They finally selected NT2 as an optimal sgRNA because it contains relatively higher ability to generate indels and lower predicted off-target potential. Our s2 exactly matches with their NT2 (in comparison, RS2 ranks this optimal sgRNA at the sub-optimal 3rd position, while CRISPRscan ranks it at the 28th position among all the potential sgRNAs for cutting Nrl). This suggests that our TSAM cleavage efficiency regression method is quite accurate for recommending good sgRNAs for disease gene editing. Our method is indeed useful to suggest only several top-ranked sgRNAs (e.g., top 3) for narrowing down the search scope in the subsequent filtering such as the off-target prediction and in vivo

experimental test. Such a recommendation approach can save time and costs, meanwhile achieving satisfactory accuracy.

De Ravin et al. (De Ravin et al. 2017) investigated a gene repair problem with CRISPR/Cas9 to cure patients with X-linked chronic granulomatous disease that arises from mutations in CYBB (C676T substitution in exon 7 of CYBB gene). Different from the above case study, to correct the point mutation, the cutting site should be close to the mutation site. Four potential sgRNAs (gRNA1, gRNA2, gRNA3 and gRNA8) whose cutting sites are near the mutation site were tested. They found that gRNA2 (5'-CACCCAGATGAATTGTACGT-3') had the maximal cutting efficiency. By our TSAM (exactly, TSAM-MT1 is used, because these sgRNAs cut at non-coding regions), the predicted scores of the four sgRNAs are: 0.310 for gRNA1, 0.693 for gRNA2, 0.534 for gRNA3 and 0.243 for gRNA8. For comparison, the predicted scores by RS2 are quite different as 0.364, 0.704, 0.555 and 0.351 respectively. On the other hand, CRISPRscan could detect just gRNA3 (score=28) and gRNA8(score=35), but not gRNA1 or gRNA2 (gRNA1 and gRNA2 start with 'TT' and 'CA' respectively, thus they cannot be expressed from the T7 promoter and predicted by CRISPRscan (Moreno-Mateos et al. 2015)). Thus, TSAM can accurately recommend the optimal sgRNA for the mutation correction case as well.

## 6.4 Conclusion

In this chapter, we propose a two-step averaging method (named TSAM) to conduct regressions on the cleavage efficiencies of sgRNAs. This work solves the research question **Q2**. The contributions of this part work has been simply concluded in Chapter 1 **Section 1.3 C4**. The following contents give some complementary descriptions of this chapter's contribution.

In our TSAM, the first-step cleavage efficiency scores are predicted by an optimized XGBoost regressor. This step also ranks the features' importance for feature selection. At the second step, an SVM regression model is

constructed using the pHMM features combined with the top-ranked features selected by the first step. The first score and the second score are averaged as the cleavage efficiency of each sgRNA in the prediction. Our regression method can be easily converted into a binary classification method for the distinction between high-efficiency sgRNAs and low-efficiency sgRNAs. TSAM was evaluated on 11 benchmark datasets containing thousands of sgRNAs editing human, mouse and zebrafish genome sequences and on additional 12 datasets of different expression system. The performance of TSAM was compared with the state-of-the-art methods to prove its superior performance. Two case studies have also demonstrated the effectiveness of TSAM.

Our future work will focus on the integration of off-target prediction methods with the current on-target efficiency prediction algorithm to build a more comprehensive tool for sgRNA design where higher efficiency and specificity can be achieved simultaneously. In addition, more definitions of 'PAM' will be considered for TSAM. The cross-species cross-expression system performance evaluation will be investigated in the near future when the supporting datasets are publicly available.

# Chapter 7

# Recognition of CRISPR/Cas9 Off-target Sites Through Ensemble Learning of Uneven Mismatch Distributions

## 7.1 Introduction

With the increasing number of investigations focusing on mechanism discovery and engineering transformation of CRISPR/Cas9, practical uses of this system for clinical applications (Yin, Zhang, Qu, Zhang, Putatunda, Xiao, Li, Xiao, Zhao, Dai et al. 2017, Roper, Tammela, Akkad, Almeqdadi, Santos, Jacks & Yilmaz 2018) or other gene editing applications (Hsu et al. 2014, Swiech et al. 2015, Kramer, Haney, Morgens, Jovičić, Couthouis, Li, Ousey, Ma, Bieri, Tsui et al. 2018, Najm, Strand, Donovan, Hegde, Sanson, Vaimberg, Sullender, Hartenian, Kalani, Fusi et al. 2018) are also widely explored. It has been introduced in Chapter 1 that CRISPR/Cas9 with a specific sgRNA can edit at the right region of its target gene (i.e., on-target editing site), meanwhile it may bind and edit at other unintended regions as well (i.e., off-target editing site see **Fig. 7.1** as an example). As

off-target editing can cause serious toxic effects, it is of great importance to design an optimal sgRNA such that it can achieve a high on-target editing efficiency but has little or no off-target editing possibilities.

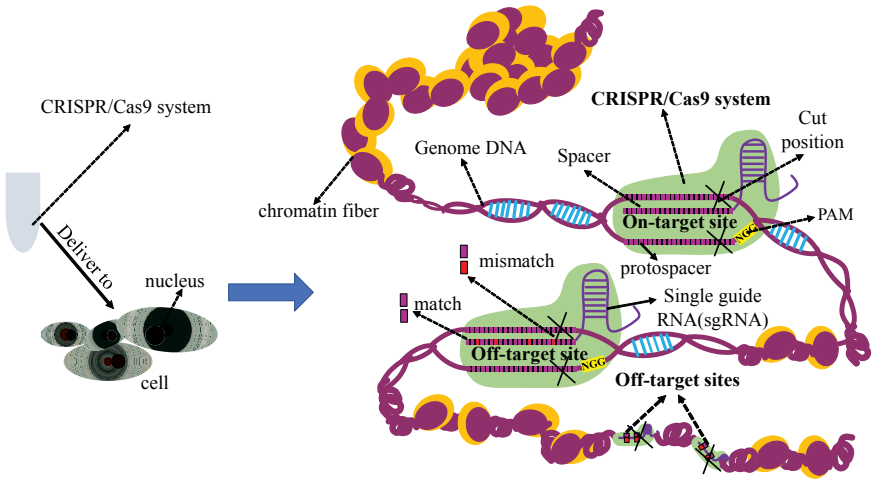


Figure 7.1: **An example of on-target site and off-target sites. The on-target site is the expected binding site for an sgRNA.** The off-target sites are unintended binding sites and the off-target editing effect should be avoided in practical use. The spacer in the sgRNA is the RNA version of the protospacer 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.

Small mismatches can be tolerated in the binding of a sgRNA with its editing site (Fu et al. 2016, Lee et al. 2016). This implies that the on-target editing site of a sgRNA and its off-target sites are homology sequences sometimes with small nucleotide differences. Therefore, off-target editing can possibly happen at any region in a genome-wide scale as long as the region contains a 3nt protospacer adjacent motif (PAM, usually 'NGG') and a 20nt protospacer sequence of minor mismatches. Such sequence

regions are all *candidate sites* of the off-target editing. As some of the candidates, maybe all of them, are not edited by the CRISPR/Cas9 system (i.e., no-editing effect), accurate recognition of the real off-target editing sites is a computationally challenging question and critically important for assessing the target specificity of the sgRNA as an optimal sgRNA. This research aims to make accurate predictions of off-target sites for a given sgRNA, assuming it has a high on-target editing efficiency. Discussions about on-target cutting efficiency prediction can be referred to those provided by CRISPR Design (Hsu et al. 2013), sgRNA Designer (Doench et al. 2016) and the method described in Chapter 6.

In Chapter 2, we have reviewed those previous efforts to address the off-target site detection problem by bioengineering or bioinformatics methods (Peng, Lin & Li 2016, Kleinstiver et al. 2016). The high-throughput sequencing methods (wet-lab technologies) include GUIDE-seq (Tsai et al. 2015), Digenome-seq (Kim et al. 2015), HTGTS (Frock et al. 2015), multiplex Digenome-seq (Kim et al. 2016), and CIRCLE-seq (Tsai et al. 2017). These approaches adopted next generation sequencing technologies to detect off-target sites at a large scale without bias, providing bona fide results. However, the experiments are always costly and time-consuming. State-of-the-art computational methods (mainly the mismatch information scoring methods) include CCTOP (Stemmer et al. 2015), MIT-score (Hsu et al. 2013), CROP-IT (Singh et al. 2015) and CFD (Doench et al. 2016), providing complementary results to the wet-lab data. These computational methods all predict cutting probabilities at the off-target sites through scoring rules. The rules are derived by analysis on the cutting efficiency changes after variations of a mismatch's position and/or mismatch type. Regions with higher scores are considered as highly-possible off-target sites. One limit of these methods is that they do not have a consistent threshold to determine whether a candidate site is a real off-target site or not. Furthermore, these rules are unable to rule out off-target sites of low cutting efficiencies but which should be also avoided in the practical use of

CRISPR/Cas9.

We propose to use sequence pairs to train a binary classification model for determining whether a given region is an off-target editing site or a no-editing site, instead of predicting the cutting efficiencies. Sequence pairs are our newly introduced concept to effectively capture integrative characteristics of the off-target editing sites of a sgRNA when combining with its on-target editing site. Let $onTSeq$ denote the sequence of the on-target site, $offTSeq$ denote an off-target editing site, and $noEdSeq$ denote a no-editing site of a sgRNA. Then $\langle onTSeq, offTSeq\rangle$ represents the sequence pair of the on-target site and the off-target site, and similarly $\langle onTSeq, noEdSeq\rangle$ represents the sequence pair of the on-target site and the no-editing site.

There exist significant differences between these on-target-off-target sequence pairs and the on-target-no-editing sequence pairs. For instance, the GC count composition change of off-target sequences is much bigger than that of no-editing site sequences when both comparing with the on-target site sequence. The mismatch distributions of these two classes of sequence pairs are also different—the 5'-end close regions contain more mismatches in the off-target sites than in the no-editing sites. Similar observations have been discussed in literature (Hsu et al. 2013, Wang et al. 2014). Furthermore, the no-editing sites' mismatches are about evenly distributed among the 20nt binding sites but the off-target sites' are not. In addition, at the first position from the 5'-end, 'G-A' mismatches are preferred in the off-target bindings while non-off-target binding likes 'G-T' more. These significant divergences between these two classes of sequence pairs provide us effective features for constructing a reliable machine learning model to make predictions of off-target sites.

The sequence pair $\langle onTSeq, offTSeq\rangle$ is called a positive sample of off-target binding, while $\langle onTSeq, noEdSeq\rangle$ is a negative sample of off-target binding. We collected positive and negative samples by going through many sgRNAs, their off-target sites and their no-editing sites for the training of our classification model. When a candidate site $canSeq$

of off-target editing is given, the classification method makes a prediction of whether $\langle onTseq, canSeq \rangle$ is a positive sample or a negative sample. If $\langle onTseq, canSeq \rangle$ is predicted as a positive sample, then $canSeq$ is an off-target editing site, otherwise, it is a no-editing site of the sgRNA. An optimal sgRNA is a sgRNA having few off-target editing sites after its all possible editing sites are screened.

In the performance test of our prediction method, we used two positive data sets. One contains off-target editing sites determined by low-throughput techniques and the other contains those determined by high-throughput techniques. The negative data set contains the genome-wide no-editing sites (allowing up to 6 mismatches). For each sequence pair from these two classes, we compute a feature vector covering the GC count characteristics and the mismatch distribution differences. Then, an ensemble support vector machine (SVM) classification model is constructed to recognize off-target sites of a test sgRNA. In a cross-dataset validation, we obtained an AUROC 0.9948 and an AUPRC 0.3323, outperforming MIT-score's 0.9807 and 0.2922, CCTOP's 0.9058 and 0.1341, CROP-IT's 0.8945 and 0.1255 and CFD's 0.8561 and 0.0453. In a further leave-one-guide-out cross-validation (logocv), our model achieved an average AUROC 0.9926 and an average AUPRC 0.4571 for 29 sgRNAs, much better than CROP-IT's 0.9160 and 0.1086, CCTOP's 0.9021 and 0.1407, CFD's 0.8835 and 0.0844 and better than MIT-score's 0.9783 and 0.2960.

Our two case studies related to the design of sgRNAs for curing retinal degeneration (Yu et al. 2017) and hearing loss (Gao, Tao, Lamas, Huang, Yeh, Pan, Hu, Hu, Thompson, Shu et al. 2018) demonstrated that our method can successfully recommend the optimal sgRNAs. Especially, in the case of curing hearing loss, our method can detect more off-target sites than the above state-of-the-art prediction methods, matching almost exactly with the off-target sites detected by a sequencing technique GUIDE-seq (Tsai et al. 2015). Our off-line tool can be downloaded from the site: https://github.com/penn-hui/OfftargetPredict.

## 7.2 Materials and Methods

### 7.2.1 Datasets for training and testing the prediction model

We collected two positive sample sets to train and test our prediction model. The first one contains on-target-off-target sequence pairs $\langle onTSeq, offTSeq\rangle$ in which the off-target sites have been experimentally validated through low-throughput techniques such as the targeted PCR and flanking PCR (Hsu et al. 2013, Cho, Kim, Kim, Kweon, Kim, Bae & Kim 2014, Kim et al. 2015, Wang, Wang, Wu, Wang, Wang, Qiu, Chang, Huang, Lin & Yee 2015, Ran, Cong, Yan, Scott, Gootenberg, Kriz, Zetsche, Shalem, Wu, Makarova et al. 2015, Kim et al. 2016). We downloaded these data from the supplementary file of (Haeussler et al. 2016). There are total 215 unique and reliable $\langle onTSeq, offTSeq\rangle$ sequence pairs related to 29 sgRNAs' on-target editing sites and their off-target editing sites. We name this positive sample set a low-throughput positive set denoted by $D_+^{low}$.

The second positive sample set consists of $\langle onTSeq, offTSeq\rangle$ sequence pairs, where the off-target sites were detected by high-throughput sequencing techniques. These techniques include GUIDE-seq (Tsai et al. 2015), Digenome-seq (Kim et al. 2015), HTGTS (Frock et al. 2015), multiplex Digenome-seq (Kim et al. 2016), and CIRCLE-seq (Tsai et al. 2017). Those off-target sites detected by at least two of these five techniques are called reliable off-target sites. This sample set is named a high-throughput positive sample set denoted by $D_+^{high}$. This data set is associated with 11 sgRNAs, a subset of the above 29 sgRNAs in $D_+^{low}$. The identical sequence pairs in $D_+^{low}$ are excluded from $D_+^{high}$. We note that among the samples obtained by Digenome-seq, those ones having been validated by targeted deep sequencing are regarded as reliable positive samples. Only the remaining samples were combined with the other four techniques' detected samples to select additional reliable positive samples. As a result, there are 527 unique and reliable sequence pairs in $D_+^{high}$. The union of $D_+^{low}$ and $D_+^{high}$ is denoted by

$D_+^{all}$. More details of these two data sets are summarized in **Table 7.1**.

In the construction of the negative sample set, we adopted an off-line tool Cas-OFFinder (Bae, Park & Kim 2014) to find genome-wide candidate editing sites *canSeq* which can have no more than 6 mismatches and contain the PAM of 'NGG' (where the mismatches in the last 2nt 'GG' are also counted) for each of the 29 sgRNAs. The latest human reference genome version hg38 was downloaded from ensembl (Aken, Ayling, Barrell, Clarke, Curwen, Fairley, Fernandez Banet, Billis, Garca Girn, Hourlier, Howe, Khri, Kokocinski, Martin, Murphy, Nag, Ruffier, Schuster, Tang, Vogel, White, Zadissa, Flicek & Searle 2016). Those candidate editing sites having been collected in positive sets were excluded in the construction of the negative sample set. There are 408260 unique negative samples. This data set is denoted by $D_-^{Cas}$. These three datasets are stored in our **Supplementary file 18**.

Table 7.1: **The datasets for constructing the positive sample sets.**

| sgRNA number | technique | sample number | literature |
|---|---|---|---|
| 4 | targeted PCR | 46 | *Hsu et al.* (2013) |
| 10 | targeted PCR | 106 | *Cho et al.* (2014) |
| 2 | PCR | 24 | *Kim et al.* (2015) |
| 2 | targeted PCR | 13 | *Wang et al.* (2015) |
| 2 | flanking PCR | 19 | *Ran et al.* (2015) |
| 10 | PCR | 21 | *Kim et al.* (2016) |
| 10 | GUIDE-seq | 403 | *Tsai et al.* (2015) |
| 10 | Digenome-seq | 248 | *Kim et al.* (2015) |
| 4 | HTGTS | 84 | *Frock et al.* (2015) |
| 11 | multiplex Digenome-seq | 954 | *Kim et al.* (2016) |
| 11 | CIRCLE-seq | 7104 | *Tsai et al.* (2017) |

### 7.2.2 Integrative characteristics of sequence pairs

We introduce GC count and mismatch distribution to describe the integrative characteristics of sequence pairs, and conduct mismatch distribution analysis such as position-specific mismatch frequency comparison and position-specific mismatch preference analysis between the positive and negative sequence pairs.

GC count of a sequence $S$ is defined as the number of guanine (G) and cytosine (C) in $S$. It is denoted as $GC\_count(S) = numG(S) + numC(S)$, where $numG(S)$ represents the number of 'G' in $S$ and $numC(S)$ is the number of 'C' in $S$. The GC count difference between sequence $S_1$ and sequence $S_2$ is the GC count of $S_2$ subtracting the GC count of $S_1$. It is denoted as $\Delta GC(S_1, S_2) = GC\_count(S_2) - GC\_count(S_1)$.

A mismatch is traditionally referred to as the base pairing at a position of a sgRNA and its DNA target site disagreeing with the rules that 'U' pairs with 'A' (U-A), 'A' pairs with 'T'(A-T) and 'G' pairs with 'C' (G-C). In this work, if the two nucleotides at the same position of an $onTSeq$ and its corresponding $offTSeq$ are different, these two nucleotides form a *mismatch*. Total 12 types of mismatches can happen, namely, $Mis = \{$'A-T', 'A-C', 'A-G', 'T-C', 'T-G', 'T-A', 'G-A', 'G-T', 'G-C', 'C-A', 'C-T', 'C-G'$\}$. On the other hand, 'A-A', 'G-G', 'C-C' and 'T-T' are called matches between $onTSeq$ and its homology $offTSeq$. See **Fig. 7.2** for an example of sequence pair $\langle onTSeq, offTSeq\rangle$ and their mismatches.

Let $t = 1, 2 \ldots,$ or 23 be a position number in $onTSeq$, then the mismatching frequency of position $t$ in $D_+^{all}$, denoted by $mfreq(t, D_+^{all})$, is computed by

$$mfreq(t, D_+^{all}) = misnum(t, D_+^{all})/misnum(D_+^{all}) \tag{7.1}$$

where $misnum(t, D_+^{all})$ counts the number of mismatches at position $t$ in $D_+^{all}$ and $misnum(D_+^{all})$ counts the number of all the mismatches in $D_+^{all}$. Similarly, we compute $mfreq(t, D_-^{Cas})$. Then we compare these two mismatching frequencies at every position between the two classes of sequence

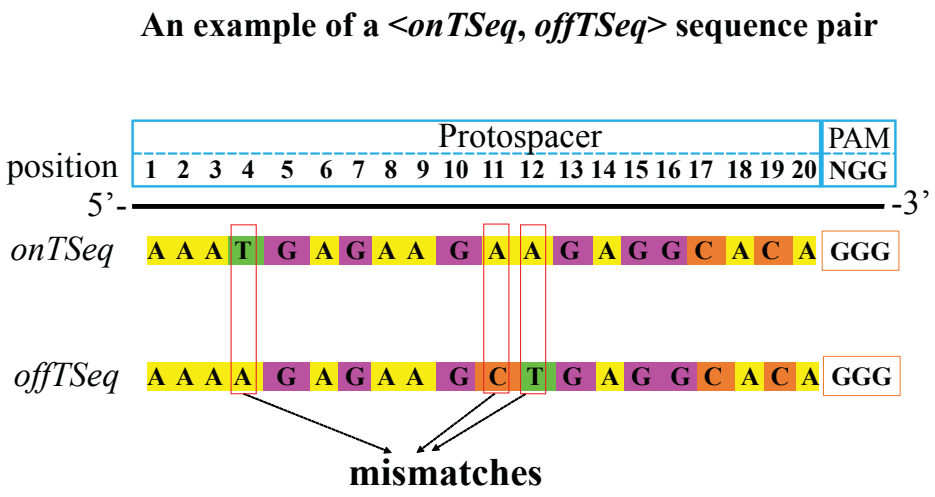


Figure 7.2: **An example of a sequence pair** $\langle onTSeq, offTSeq \rangle$**.** The mismatches are those pairs of nucleotides at the given position but with different nucleotide type such as at the positions of 4, 11 and 12

pairs.

Let $mis \in Mis$ be a given type of mismatch at position $t$, then the frequency of $mis$ at $t$ in $D_+^{all}$, denoted by $mfreq(mis, t, D_+^{all})$, is defined as:

$$mfreq(mis, t, D_+^{all}) = misnum(mis, t, D_+^{all})/misnum(t, D_+^{all}) \tag{7.2}$$

where $misnum(mis, t, D_+^{all})$ counts the number of a given $mis$ at position $t$ in $D_+^{all}$, and $misnum(t, D_+^{all})$ counts all 12 types of mismatches at $t$ in $D_+^{all}$. Such position-specific mismatch preference in the negative sequence pair data set, $mfreq(mis, t, D_-^{Cas})$, can be similarly defined.

The position-specific mismatching frequency comparison between two classes of sequence pairs is through $mfreq(t, D)$, while the position-specific mismatch preference analysis is based on $mfreq(mis, t, D)$. Here, the Two-sample Kolmogorov-Smirnov test (Two-sample K-S test) (Lilliefors 1967) is used with a threshold of p-value<=0.05 to evaluate whether the differences are significant. To get rid of the sample size's effect on the mismatch

distribution comparison, we also randomly select 100 subgroups of negative sequence pairs to compare with the positive sequence pairs. Each of the negative subgroups has the same size as the positive group.

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

These GC counts characteristics and mismatch preferences are exploited as new features, and we convert every sequence pair into a vector under the new feature space. The new feature space consists of two subsets of features. The first subset covers the nucleotide composition change features; the second subset covers the position-specific binary mismatch features. Let an on-target sequence denoted as $onTSeq = s_1 s_2, \ldots, s_i, \ldots, s_{23}$, and an off-target sequence denoted as $offTSeq = s_1 s_2, \ldots, s_j, \ldots, s_{23}$, where $s_i, s_j \in \{A, G, C, T\}$, the first 20nt sequences represent the protospacer sequences, and the last 3nt sequences are the PAM sequences. Then $\langle onTseq, offTSeq\rangle$ is converted into a new feature vector by the following definitions and steps.

#### Nucleotide composition change features

The nucleotide composition change features (NCC) are: $\Delta GC$ (simply denoted as $f1$), GC percent change, GC skew change, AT skew change and the change of the ratio between GC skew and AT skew.

- **GC percent change ($f2$).** The GC percent of a sequence $S$, denoted $GC\_per(S)$, is defined as $GC\_count(S)/length(S)$, where $length(S)$ stands for the number of nucleotides in $S$. The GC percent change from sequence $S_1$ to sequence $S_2$ is defined as $GC\_per(S_2) - GC\_per(S_1)$.

- **GC skew change ($f3$):** GC skew change is a measure of the distribution of guanine (G) and cytosine (C) across the two DNA strands $S_1$ and $S_2$ (Ginno, Lim, Lott, Korf & Chédin 2013). As reported (Jiang, Taylor, Chen, Kornfeld, Zhou, Thompson, Nogales &

Doudna 2016), GC skew is one of the key factors predisposing to R-loop formation and R-loop formation is a necessary step for CRISPR/Cas9 system to cut its target site. The GC skew feature is adopted to characterize the sequences. GC skew of sequence $S$ is computed as $GC\_sk(S) = (numG(S) - numC(S))/GC\_count(S)$, and GC skew change from sequence $S_1$ to sequence $S_2$ is $GC\_sk(S_2) - GC\_sk(S_1)$.

- **AT skew change ($f4$):** Similarly, we also compute the AT skew as a feature for a given sequence $S$. That is, $AT\_sk(S) = (numA(S) - numT(S))/(numA(S) + numT(S))$. Then AT skew change from sequence $S_1$ to sequence $S_2$ is $AT\_sk(S_2) - AT\_sk(S_1)$.
- **Change of ratio of $GC\_sk$ and $AT\_sk$ ($f5$):** The ratio of $GC\_sk$ and $AT\_sk$ for a sequence $S$ is defined as $R\_sk(S) = GC\_sk(S)/AT\_sk(S)$, and the ratio change from sequence $S_1$ to sequence $S_2$ is $R\_sk(S_2) - R\_sk(S_1)$.

The first subset of new features converted from a sequence pair $\langle S_1, S_2\rangle$ in $D_+^{all}$ or in $D_-^{Cas}$ is $<f1, f2, f3, f4, f5>$. For any $\langle onTSeq, canSeq\rangle$, it is similarly converted.

### Position-specific binary mismatch features

For a pair of sequences $S_1$ and $S_2$, between the $t$-th position of $S_1$ and $S_2$, there are four types of nucleotide matches (i.e., 'A-A', 'G-G', 'C-C', and 'T-T') and there are 12 types of nucleotide mismatches (i.e., $mis \in Mis$). The matching or mismatching status of each position between $S_1$ and $S_2$ can be converted into a 12-dimension binary vector. If the position is matched between $S_1$ and $S_2$, then this position is converted into an all-zero 12-dimension vector $<0, 0, \ldots, 0>$. If the position is mismatched as the $i$-th mismatch type, then this position is converted into a binary vector $<0, \ldots, p_i, 0, \ldots, 0>$, where only $p_i$ is 1. Merging all these 12-dimension vectors, the sequence pair $\langle S_1, S_2\rangle$ can be converted into a $(12*23)$-dimension vector. We convert every sequence pair in $D_+^{all}$ and $D_-^{Cas}$ by this way. These

position-specific binary features can characterize the mismatch distributions of the sequence pairs.

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

Merging the 5-dimension nucleotide composition change-related feature vector and the 276-dimension mismatch-related feature vector, every sequence pair in this study is converted into a 281-dimension feature vector. The positive and negative 281-dimension vectors are used to train a machine learning method for the prediction of whether a candidate site is an off-target editing site or a no-editing site of an sgRNA.

We propose to use an ensemble SVM classifier to predict off-target sites. The motivation is that the collected datasets are extremely imbalanced, and ensemble learning is a good strategy to improve the prediction accuracy and stability. It was also reported that random under-sampling is one of the effective strategies for addressing imbalanced learning problem (He & Garcia 2009). Thus, we randomly select the same number of negative samples as that of the positive samples to train base classifiers for $n$ times. An ensemble SVM classifier is built by averaging the $n$ base classifiers' output probabilities.

The construction of our prediction model has two procedures: the optimization step and the evaluation step. In the optimization step, we optimize three super-parameters: penalty parameter C and the parameter gamma of RBF kernel for SVM (Libsvm v3.22 (Chang & Lin 2011)) and the ensemble size $n$, by a leave-one-guide-out cross-validation (logocv). The training dataset for the logocv composes of $D_{+}^{high}$ and those negative samples in $D_{-}^{Cas}$ corresponding to the involved 11 sgRNAs. During the logocv, samples corresponding to each sgRNA are adopted as validation data in turn and the remaining samples are used as training data. Achieving the best AUROC is used as the criteria to determine the optimal parameters.

We evaluate and compare our method with the state-of-the-art methods CCTOP (Stemmer et al. 2015), MIT-score (Hsu et al. 2013), CROP-IT

(Singh et al. 2015) and CFD (Doench et al. 2016) by a cross-dataset validation and a logocv. The cross-dataset validation is conducted by training the classifier with the above training dataset and testing it with the samples related to the remaining 18 sgRNAs in $D_{+}^{low}$ that have been excluded from the training dataset. The logocv is implemented on the dataset combining $D_{+}^{all}$ and $D_{-}^{Cas}$.

The AUROC (Aera Under Receiver Operating Characteristic curve) and the AUPRC (Aera Under Precision-Recall curve) are adopted as the performance indexes to show how different methods can rank the positive samples comparing with those negative ones. ROC curve and PR curve are popular visual representation tools for illustrating a classifier's performance and their corresponding AUROC and AUPRC are used to quantify the classifier's performance (He & Garcia 2009). Especially, the PR curve was thought to be a more informative representation of performance assessment under highly imbalanced data (Davis & Goadrich 2006, He & Garcia 2009).

We note that the four existing methods have taken a scoring strategy but not the machine learning approach. These traditional methods generate scoring rules converted from correlations between the mismatch numbers, positions, and cutting efficiency changes among huge amount of simulative off-target bindings in their own datasets (Hsu et al. 2013, Doench et al. 2016). This makes their scoring functions hardly adaptable to our collected datasets because our positive sample size is much smaller than theirs and we do not have exact cutting efficiency change values. It is also impossible for us to train our models on their datasets as there is no threshold used for labeling their samples or the datasets are unaccessible. Thus, in the performance comparison, we use their already well-tuned scoring rules instead of re-training them on our data.

## 7.3 Results

We first report GC composition related characteristics and mismatch enrichment and preferences in the comparison between on-target-off-target sequence pairs and on-target-no-editing sequence pairs. Then, we report the superior prediction performance of our method in comparison with the state-of-the-art computational methods. We also present how the predicted off-targets by the computational methods overlap with those detected by the high-throughput sequencing techniques. At last, two case studies of applying our method to assist optimal sgRNA selection for disease treatment are described.

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

For the on-target-off-target sequence pairs in $D_+^{all}$, on average the GC count decreases more than the on-target-no-editing sequence pairs in $D_-^{Cas}$ does. The positive samples have a mean $\Delta GC$ = -1.09, while the negative samples have a mean $\Delta GC$ = -0.71. This difference is significant with p-value=2.31E-07 by the two-sample K-S test (Lilliefors 1967). For the randomly selected 100 subgroups of negative samples comparing with the positive samples (Section 7.2.2 last paragraph), all of them had significant differences (p-value<0.05); and all of the randomly selected negative data sets have smaller drop of the GC count than the positive samples.

The position-specific mismatching frequencies in the positive samples (i.e., $mFreq(t, D_+^{all})$) are unevenly distributed over the positions $t = 1, 2, \ldots, 20$, peaking at the 5'-end close positions. However, in the negative samples, the mismatches seem to be uniformly distributed, all having about 5% of the mismatches (i.e, $mFreq(t, D_-^{Cas})$ =5% for all $t$). See an illustration of these mismatching frequency distributions at **Fig 7.3**. The mismatching frequencies at the 20 positions are significantly different between the positive and negative samples (p-value=0.0082). This suggests that if a candidate editing site of a sgRNA has mismatches with the on-target site accumulating

at the 5' end, this candidate site is more likely to be an off-target editing site instead of no-editing site.

100 rounds of similar comparison analyses were conducted for the set of positive samples and an equal-size of negative samples randomly selected from the entire set of negative samples (Section 7.2.2 last paragraph). About 98% of these comparisons showed significant differences (p-value<0.05). Especially, the 5'-end adjacent two positions contains more mismatches than the other regions (12% and 9% at the 1st and the 2nd position vs. no more than 7% at the other positions except for the 8th position ordered from 5' to 3'). Similar phenomena have been previously reported by (Hsu et al. 2013, Fu et al. 2013, Pattanayak, Lin, Guilinger, Ma, Doudna & Liu 2013). Thus the 5'-end nucleotides are not as conserved as the nucleotides in the other regions if they can be edited by sgRNAs. This observation is consistent with the literature comment that the 5'-end truncated sgRNAs can decrease the level of off-target potentials (Ren, Yang, Xu, Sun, Mao, Hu, Yang, Qiao, Wang, Hu et al. 2014, Sternberg & Doudna 2015, Kleinstiver et al. 2016).

From the analysis of mismatch type frequency distributions at given positions $t = 1, 2, \ldots, 20$, we found that there exists strong mismatch type preference. For example, the position immediately adjacent to 5'-end (the first position from 5' to 3') has significant differences among the 12 types of mismatches (p-value=0.0046). Though both the positive samples and negative samples contain more 'G-A', 'G-T' and 'G-C' mismatches at the first position (this may due to most of the first nucleotide of the spacers are 'G'), the positive samples prefer the 'G-A' mismatch ($mfreq(\text{'G-A'}, 1, D_+^{all}) = 0.4353$) over the 'G-T' mismatch ($mfreq(\text{'G-T'}, 1, D_+^{all}) = 0.224$) or the 'G-C' mismatch ($mfreq(\text{'G-C'}, 1, D_+^{all}) = 0.2965$). On the other hand, there are relatively more 'G-T' mismatches at position 1 in the negative samples than 'G-A' or 'G-C' ($mfreq(\text{'G-T'}, 1, D_-^{Cas}) = 0.2525$, $mfreq(\text{'G-A'}, 1, D_-^{Cas}) = 0.2279$ and $mfreq(\text{'G-C'}, 1, D_-^{Cas}) = 0.1616$). The 100 balanced comparisons also show significant differences at this position. More details about this comparison can be found in our **Supplementary file 19**.

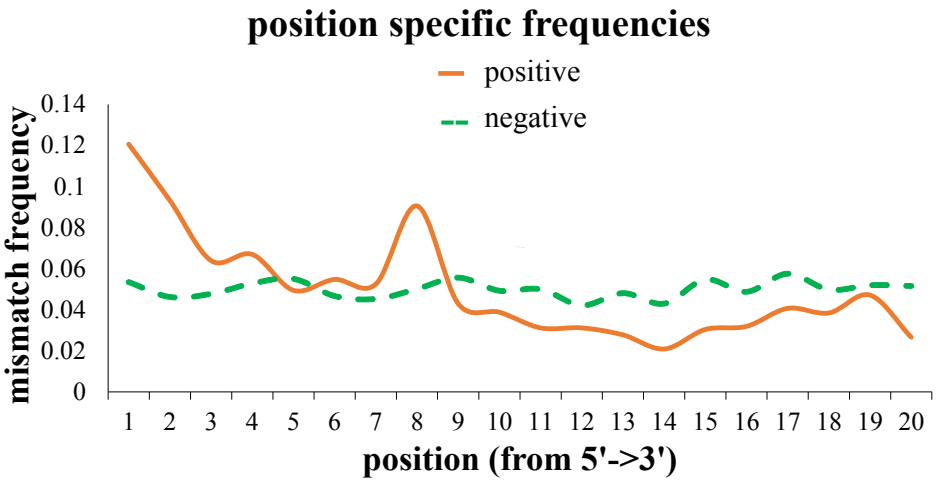


Figure 7.3: **Comparison of the mismatch distributions in the positive and negative sample sets.** The lines depict the remarkable distribution differences between the two groups.

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

Our training dataset containing the 11 sgRNAs related samples from $D_{+}^{high}$ and $D_{-}^{Cas}$ was applied to optimize the parameters such as the penalty parameter C ($c = 2^C$) and the parameter gamma ($g = 2^G$) for RBF kernel of the SVM, and the ensemble size $n$. During each fold of the logocv on this training dataset, we optimized $c$ and $g$ with a grid search method where $C, G \in [-6 : 1 : 6]$ and set $n = size(negative)/size(positive)$, where $size(negative)$ stands for the number of negative samples in the training data. When $C$=1 and $G$=-4, the highest average AUROC=0.9819 was achieved. Fixing $C$=1 and $G$=-4, we explored how ensemble size $n$ affects the prediction performance. We found that a bigger $n$ can indeed decrease the standard deviation of the prediction performance, however, the AUROC just changes no more than 0.01 and the running time increases rapidly. Thus, $n = 40$

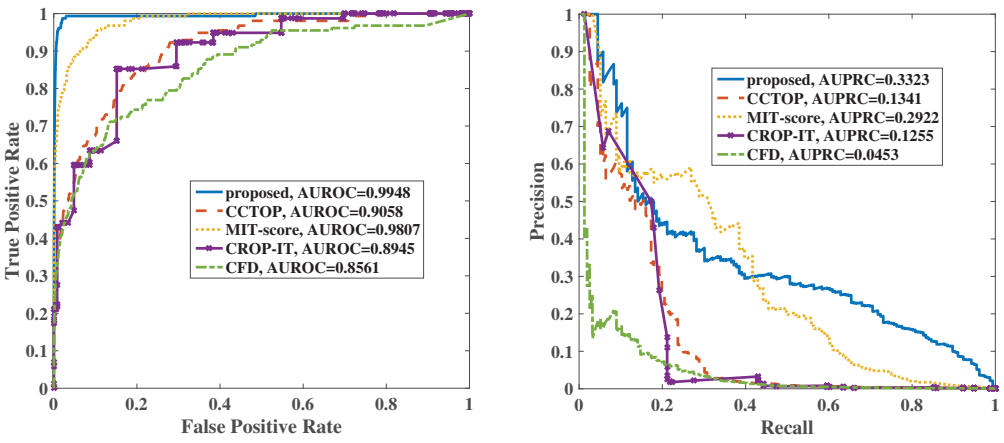


Figure 7.4: **Receiver Operating Characteristic curves (left) and Precision-Recall curves (right) for the cross-dataset validation of our proposed method and the four state-of-the-art methods.**

was selected at last (see **Supplementary file 19** to find more details of the parameter selection).

In the cross-dataset validation, we trained our ensemble model with the above optimal parameters and on the whole *training* dataset. The performances of our proposed method and the four state-of-the-art methods were tested on the dataset containing the *remaining* 18 sgRNAs related sequence pairs in $D_+^{low}$ and $D_-^{Cas}$. The ROC curves and the Precision-Recall curves are shown in **Fig 7.4** and the detailed performance statistics are listed in **Table 7.2** (the second and third column).

These curves and the detailed performance measurements clearly suggest that both our proposed method and MIT-score can achieve much better performance than the other three methods (under the cross-dataset test). Furthermore, our proposed method also works better than MIT-score. This implies that under the same false positive rate or recall, our method can obtain the best true positive rate and precision.

The logocv performance by the five methods was achieved on the dataset $D_+^{all}$ merging $D_-^{Cas}$ which contains the all 29 sgRNAs' on-target sites and

Table 7.2: **AUROC and AUPRC scores of the proposed method and the state-of-the-art methods in various tests.**

| Methods | cross-dataset validation | | logocv[1] | |
|---|---|---|---|---|
| | AUROC | AUPRC | AUROC | AUPRC |
| proposed | 0.9948 | 0.3323 | 0.9926 | 0.4571 |
| CCTOP | 0.9058 | 0.1341 | 0.9021 | 0.1407 |
| MIT-score | 0.9807 | 0.2922 | 0.9783 | 0.2960 |
| CROP-IT | 0.8945 | 0.1255 | 0.9160 | 0.1086 |
| CFD | 0.8561 | 0.0453 | 0.8835 | 0.0844 |

[1] leave-one-guide-out cross-validation on combined dataset

off-target sites. Thus, this logocv had 29 rounds each corresponding to one sgRNA. In each round, all sequence pairs related to the same one sgRNA were reserved as test data, while the remaining sequence pairs were used to train the ensemble prediction model. The mean AUROC and mean AUPRC over the 29 rounds of tests are listed in the last two columns of **Table 7.2**. Our proposed method outperforms the four existing methods on both AUROC and AUPRC measurements. MIT-score works the best among the four state-of-the-art methods; however, our proposed method still exceeds its performance by a 0.1611 AUPRC score and a 0.0143 AUROC score. The comparison between our method and the two most recently published methods such as CRISTA (Abadi et al. 2017) and Elevation (Listgarten et al. 2018) can be found in our **Supplementary file 19**.

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

We carried out analysis to understand how our predicted off-target sites overlap with those determined by high-throughput sequencing techniques. Recently developed high-throughput sequencing techniques include GUIDE-

seq (Tsai et al. 2015), Digenome-seq (Kim et al. 2015), HTGTS (Frock et al. 2015), multiplex Digenome-seq (Kim et al. 2016), and CIRCLE-seq (Tsai et al. 2017). We compared the list of the off-target sites predicted by our method with the list of off-target sites predicted by each of these sequencing techniques.

Two sgRNAs and their on-target sites were used in this analysis. One is the sgRNA which targets to the EMX1 site (protospacer+PAM: GAGTCCGAGCAGAAGAAGAAGGG). This sgRNA is the only sgRNA whose off-target sites were sequenced by all of the five high-throughput sequencing methods. The second is the sgRNA which targets to the HEK4 site (protospacer+PAM: GGCACTGCGGCTGGAGGTGGGGG). Four of the sequencing methods (no HTGTS) had been applied in the literature to detect the off-target sites of HEK4. We note that these sequencing methods had produced different lists of off-target sites. For EMX1, there are total 15835 potential off-target sites. CIRCLE-seq, Digenome-seq, GUIDE-seq, HTGTS, and multiplex Digenome-seq detected 176, 27, 15, 13, and 142 off-target sites respectively. Some of these off-target sites were detected more than twice, the union of these off-target sites contains only 259 off-target sites (we call them the 'Integrated' detections). For HEK4, 30175 potential off-target sites were found genome-wide. CIRCLE-seq, Digenome-seq, GUIDE-seq, multiplex Digenome-seq had produced 980, 38, 133, and 215 off-target sites respectively. The union of these off-target sites contains 1011 unique ones.

In the prediction of EMX1 off-target sites by our method, the model was trained on the positive and negative data sets $D_+^{all}$ and $D_-^{Cas}$ after all sequence pairs containing the on-target site EMX1 were removed. Similarly, all sequence pairs containing the on-target site HEK4 were removed from the training data in the prediction of HEK4 off-target sites.

An overlap rate (OR) of two lists of off-target sites is used to weight how a computational method's predictions overlap with a high-throughput

sequencing method's detections:

$$OR(comM, seqM) = \frac{off(comM) \cap off(seqM)}{off(seqM)} \times 100\% \qquad (7.3)$$

where $comM$ stands for a computational method, $seqM$ represents a sequencing method, $off(comM)$ is the list of off-target sites predicted by $comM$, and $off(seqM)$ is the list of off-target sites detected by $seqM$. The four state-of-the-art computational methods were compared with our computational method to see which one is better to overlap with the sequencing methods. Because of no thresholds were provided by these literature computational methods, we selected top-ranked $N$ off-target sites according to their scores, where $N$ is the number of test samples labeled as positive by our method. The overlap rates with regard to different combinations of the computational methods and sequencing methods are depicted in **Fig 7.5**.

For the EMX1 site, our computational method predicted 673 off-target sites. The ORs are 61% (108 out of 176), 89% (24 out of 27), 100% (15 out of 15), 100% (13 out of 13), 42% (59 out of 142) comparing with CIRCLE-seq, Digenome-seq, GUIDE-seq, HTGTS, and multiplex Digenome-seq respectively, and the ORs for the 'Integrated' is 43% (112 out of 259). For the HEK4 site, we predicted 1202 off-target sites and the ORs are: 43% (417 out of 980), 92% (35 out of 38), 90% (120 out of 133), 76% (163 out of 215) and 42% (421 out of 1011) for CIRCLE-seq, Digenome-seq, GUIDE-seq, multiplex Digenome-seq, and Integrated, respectively. From **Fig 7.5**, we can see that our ensemble model predicted off-target sites overlap with those sequencing methods better than the other four computational methods. As these sequencing methods detect different and far-incomplete lists of off-target sites, we also draw the conclusion that our method can predict more complete lists of off-target sites than any of the sequencing methods or their union. Our computational tool can predict off-target sites that overlap well with the sequencing methods, thus it can be used as a supplementary tool for selecting sgRNAs with higher specificities to be further validated by sequencing techniques, for saving time and cost.

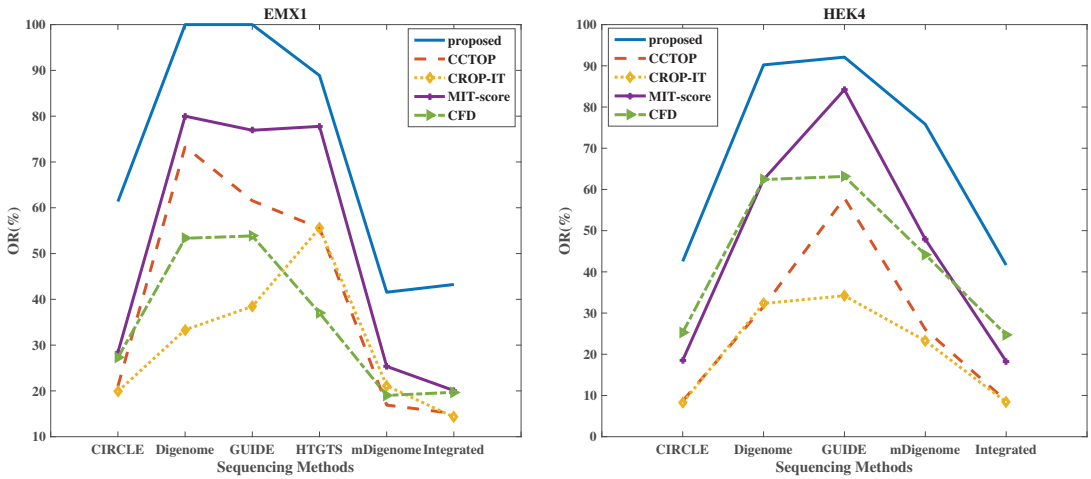


Figure 7.5: **Overlap rates of different computation methods relative to the high-throughput sequencing base methods.** The proposed method detected off-targets overlaps better than other computational methods relative to all the sequencing methods' results. Sequencing methods CIRCLE, Digenome, GUIDE, 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.

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

Off-target effect is one of the main problems in the application of CRISPR/Cas9 to cure genetic diseases. Here, we present two case studies to demonstrate how our tool can help select the optimal sgRNAs with off-target effect as less as possible. The first case study is about the application of CRISPR/Cas9 to knockdown mouse Nrl gene for preventing retinal degeneration (Yu et al. 2017). Five potential sgRNAs containing protospacer sequences NT1 to NT5 were initially designed by (Yu et al. 2017). Their on-target cutting efficiencies were estimated by in vitro experiments. These sgRNAs' possible off-target sites were predicted and combined with their cutting efficiencies to provide a selection guidance.

Table 7.3: **The ranks of the sgRNAs by considering both of their cutting efficiencies and off-target potentials.**

| sgRNA | literature | | | proposed | | | CRISPR Design | | | sgRNA Designer | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Indel(%) | ER[1] | FR[2] | otN[3] | otR[4] | FR | otN | otR | FR | otN | otR | FR |
| NT1 | 21.9 | 4 | - | 264 | 5 | 5 | 101 | 2 | 2 | - | 1 | 3 |
| **NT2** | 22.7 | 2 | **1** | 83 | 1 | **1** | 69 | 1 | 1 | - | 2 | **1** |
| NT3 | 22.5 | 3 | - | 139 | 4 | 3 | 159 | 4 | 4 | - | 5 | 4 |
| NT4 | 23.2 | 1 | - | 119 | 3 | 2 | 146 | 5 | 5 | - | 3 | **1** |
| NT5 | 18.3 | 5 | - | 95 | 2 | 3 | 115 | 3 | 3 | - | 4 | 5 |
| Tmc1-mut1 | 4.1 | 2 | - | 613 | 3 | 3 | 337 | 3 | 3 | - | 1 | **1** |
| Tmc1-mut2 | 0.74 | 3 | - | 183 | 1 | 2 | 318 | 2 | 2 | - | 3 | 2 |
| **Tmc1-mut3** | 10 | 1 | **1** | 247 | 2 | **1** | 197 | 1 | **1** | - | 2 | 3 |

[1] efficiency rank

[2] final rank

[3] off-target site number

[4] off-target site rank

We applied our method and two other web-tools, CRISPR Design (Hsu et al. 2013) (http://crispr.mit.edu/, the off-target prediction method is the previous MIT-score) and sgRNA Designer (Doench et al. 2016) (https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design, it uses the previous CFD to predict the off-target sites), to rank the five sgRNAs' off-target sites. The mouse reference genome version mm10 was downloaded from ensembl (Aken et al. 2016). As our method's input, the candidate off-target sequences of sgRNAs NT1 to NT5 were extracted by Cas-OFFinder allowing 6 mismatches at most and with the PAM 'NGG'. A sgRNA having more off-target sites is ranked lower. For a fair comparison, the cutting efficiencies were ranked according to the estimated cutting efficiencies by (Yu et al. 2017). The final rank of a sgRNA is determined by the average rank of its cutting efficiency rank (efficiency rank) and its off-target rank (ot rank). In fact, we can adjust the weights of the two kinds of ranks in practical usages. The best final rank suggests the corresponding sgRNA works the best and should be selected. The detailed results are shown in the first five rows of **Table 7.3**.

The authors (Yu et al. 2017) reported that NT2 was the optimal sgRNA as it contains better cutting efficiency and relatively lower off-target potential.

Our method and CRISPR Design mark NT2 as the best choice. However, sgRNA Designer ranks NT1 as the one with the lowest off-target potential and both of NT2 and NT4 have the final rank of 1 which may confuse the users.

The second case study is about a recent application of CRISPR/Cas9 to treat autosomal dominant hearing loss (Gao et al. 2018). Four sgRNAs were designed at first with the protospacer sequences of Tmc1-mut1 to Tmc1-mut4. We did not consider Tmc1-mut4 as it is a truncated protospacer (Our method and the other two web tools cannot support this type of spacers—20nt sequences must be required). The authors (Gao et al. 2018) tested the cutting efficiencies of these sgRNAs. Then GUIDE-seq was adopted to estimate the sgRNAs' off-target potential. Similar to the first case study, we considered both the on-target editing efficiency ranks and off-target potentials produced by our method, CRISPR Design and sgRNA Designer for the final ranking of these sgRNAs. The results are listed in the last three rows of **Table 7.3**.

Our method and CRISPR Design can both recommend the sgRNA containing Tmc1-mut3 as the optimal sgRNA, consistent with the optimal sgRNA used by Gao et al. (Gao et al. 2018). On the other hand, sgRNA Designer ranked Tmc1-mut1 as the best. Gao et al. (Gao et al. 2018) detected 10 off-target sites for the sgRNA containing Tmc1-mut3 with up to 6 mismatches in the protospacer region. We compared the off-target sites detected by our method, those by the MIT-score and those by CFD with these 10 GUIDE-seq detected ones. We found that 9 out of the 10 sites were predicted by our method (among total 247 predicted positive off-targets). In addition, all these 9 off-target sites were ranked at the top 30, where six out of them were ranked at the top 10. In comparison, MIT-score only found 8 of these GUIDE-seq validated off-target sites if we defined their top-ranked 247 ones as positive. Among these 8 sites, 4 of them were ranked at top 10. The CFD ranked 7 out of the 10 GUIDE-seq validated off-target sites at the top 247. However, only 1 of them was ranked at top 10. More details about

the comparison can be found in our **Supplementary file 19**. These two case studies partly prove that our method can effectively help sgRNA design for practical use.

## 7.4 Conclusion

In this chapter, an ensemble machine learning method is proposed for the prediction of off-target sites of sgRNAs. This method settles the research question **Q3**. Its contributions have been described in **Section 1.3 C5**. This method is based on the observation that there exist significant GC count differences and mismatch preferences between the positive on-target-off-target sequence pairs and those negative ones. Our method not only takes advantage of the information from known off-target sites but also adopts the information from those no-editing target sites. It improves the performance of off-target site prediction in comparison with the state-of-the-art computational methods; and can detect more off-target sites consistent with the bona fide detections by high-throughput sequencing methods. As demonstrated in the two case studies, our method is effective for selecting the optimal sgRNAs to treat some genetic diseases.

Our future work will focus on two areas. One area is about data collection. We will investigate which positive and negative samples are more reliable, especially for the negative samples. In addition, the samples containing bulges (Abadi et al. 2017) should also be included when abundant data are available. The second area is about the new feature space in the conversion of the sequence pairs into the new vectors. Other informative features such as cutting positions and the dinucleotide mismatch distribution can be exploited to expand the current feature space. Furthermore, a tool integrating the sgRNA on-target cutting efficiency prediction and our off-target site prediction is worth of construction for providing more comprehensive guidance for selecting the optimal sgRNAs.

# Chapter 8

# Conclusions and Future Work

## 8.1 Conclusions

The work in this thesis mainly addresses two significant bioinformatics issues namely the disease-ncRNA association prediction and the optimal design of a CRISPR/Cas9 system for gene editing. The proposed methods for solving these two issues were discussed in Chapters 3-7 and have been presented in four of my published journal papers (see the **List of Publications**). The Chapters 3-5 described the methods for disease-miRNA and disease-lncRNA association prediction while the Chapters 6&7 introduced our two machine learning methods for CRISPR/Cas9 on-target cutting efficiency prediction and off-target site detection respectively. The work and contributions of this thesis are concluded below.

In Chapter 3, a precomputed kernel matrix SVM method was introduced to predict disease related miRNAs. This method has some advantages comparing to the existing methods. Firstly, it selected those miRNA-disease pairs that the miRNAs are not significantly differentially expressed in the disease samples as reliable negative data. In this way, the binary classification of disease-miRNA pairs is possible. Secondly, the disease-miRNA pairs were represented by precomputed kernel matrices which are used as the inputs of SVM. This avoids the difficulty of mathematical representation of diseases.

In addition, it is possible to predict new associations for a given disease even if there is no already known miRNAs associating with it (from the miRNA side also works). Various tests for evaluation and comparison proved the excellent performance of this method.

The purpose of our study in Chapter 4 is twofold: proving the usefulness of our proposed precomputed kernel matrix SVM for predicting disease-miRNAs and investigating the roles of co-functional miRNA pairs in multi-diseases' development. A prioritization method was designed for selecting reliable multi-cancer related co-functional miRNA pairs from the reconstructed cancer-gene-miRNA tripartite. It applies three kinds of information: miRNA function relationship; miRNA regulation relationship in different diseases; and co-functional miRNAs' co-dysregulation relationship. Some valuable multi-cancer related co-functional miRNA pairs were obtained such as the miR-15b-miR-195. Through the gene ontology and pathway enrichment analysis, several of these pairs were proved to be really important in multi-cancers' development. According to the further comparison of the cancer-gene-miRNA and non-cancer disease-gene-miRNA networks' analysis, two conclusions can be reached: the co-function phenomenon is not unusual and the regulation of miRNAs for the development of cancers is more complex and have more unique properties than those of non-cancer diseases.

Chapter 5 focuses on the problem of predicting disease-related lncRNAs. A bagging SVM-based positive-unlabeled learning method was leveraged to settle this matter. There are at least three novelties in this study. First of all, a novel disease vectorization method was proposed. It characterizes a disease with its related genes' chromosome substructure and pathway groups' distribution properties. Secondly, a disease-lncRNA association prediction problem was transferred to be a disease-lncRNA pair classification issue. Those disease-lncRNA pairs were represented as novel feature vectors for helping the prediction of disease-lncRNA associations. The last point is that the bagging SVM was adopted to implement positive-unlabeled learning for the prediction of reliable disease related lncRNAs. In this way, the problem

of lacking reliable negative samples is solved. Through various evaluations, comparisons and case studies, the proposed method's reliability and accuracy in disease-lncRNA association prediction were confirmed.

The CRISPR/Cas9 system on-target cutting efficiency prediction issue was addressed in Chapter 6. A Two-Step Averaging Method was used to complete this task. It applied the profiled hidden Markov properties of the sgRNA sequences as novel features which capture the global characteristics of sgRNAs. Those important features selected by the XGBoost regression model were combined with the novel profiled hidden Markov properties as the input of a SVM for an enhanced prediction. The strategy of averaging the XGBoost regression results and the second step SVM's predictions as the final scores was proved to outperform the single regressions and those state-of-the-art methods. In addition, we found that those highly-active sgRNAs have lower melting temperature at the middle of the spacer, prefer to cut at 5'-end closer parts of the gene and contain more 'A' but less 'G', comparing to the low-active ones. Our further analysis also confirmed that those sgRNAs expressed from different expression systems such as the T7 promoter or a U6 promoter have inconsistent properties. Their efficiencies should be predicted with different well-trained models.

In Chapter 7, an ensemble learning method was presented to detect CRISPR/Cas9 system off-target editing sites. The off-target site detection problem was turned to be a binary classification issue by defining the sample as an on-target-candidate-target site sequence pair. A sample is labeled as positive if the candidate-target site is a real off-target site, otherwise negative. The second contribution is that the samples were characterized by the mismatch distribution properties and nucleotide composition change features. These effective characteristics were applied due to their significant differences between those positive and negative samples. For example, mismatches prefer to exist at the 5'-end closer regions of the off-target site sequences. Lastly, the ensemble learning strategy was adopted to make good use of the large amounts of negative samples and to improve the prediction

accuracy and stability. The performance tests and case studies verified that the proposed method outperforms the existing computational methods and its detected off-target sites overlap well with the wet-lab technologies' bona fide detections.

## 8.2 Future Work

Increasing studies are focusing on the disease-ncRNA gene association investigation and gene editing optimization. However, many new issues need to be addressed in these fields. Especially, the NGS technologies heavily accelerate the novel disease gene finding and annotation. The development of the bioengineering and biotechnology also impels the gene editing to its practical usage. Under this background and trends, our future work will pay attention to the following areas:

- **Novel disease related non-coding RNA gene finding**

  Non-coding RNA genes have attracted increasing attention for their significant roles in disease development. Researchers are extremely interested in lncRNA genes because of their complex functions. One of my future research topics is to find novel lncRNA genes that related to diseases especially cancers by combining the NGS data analysis with machine learning algorithms. The NGS data can provide abundant information such as the sequence, chromatin, expression level, methylation status, single nucleotide polymorphism (SNP) and tissue-specific properties. Machine learning can be applied to extract patterns from those known non-coding genes for assisting novel gene finding.

- **Non-coding RNA gene function annotation and disease mechanism investigation**

  Large scale of novel ncRNA genes have been found by adopting the RNA-seq technologies. However, their functions and exact roles in

disease development remain largely unknown. The rule that similar structure determines the similar function makes the in silico prediction of novel genes' function possible. However, this rule only works for those homology genes but not the non-homology genes. Thus, various properties should be integrated to solve this problem, e.g. regulation information, cellular location, sequence and higher level structural similarities, network features and so on. Machine learning algorithms may be applied to do the following automatically annotation jobs.

- **Systematical analysis of the gene regulation network and its application**

  The genes including ncRNA genes always form functional modules to involve in different biological activities. The coding gene's expression is regulated by various regulators such as ncRNAs. Consideration of the whole regulation network to find special regulators as the drug targets can increase drug efficiency but decrease the side effects. This systematical analysis of the regulation network can also benefit the understanding of disease occurrence mechanisms, which helps disease treatment and diagnosis.

- **Multi-disease gene editing optimization**

  A disease especially the cancer often relates to abundant genes. The editing of a single gene with CRISPR/Cas9 may cannot achieve expected treatment outcomes. Multi-disease gene editing may be an effective strategy for overcoming this limitation. In addition, the multi-disease gene editing can contribute to the gene cooperation related studies. Construction of an accurate disease model also requires the multi-disease gene editing technology. As most of the gene-editing optimization tools are for single genes, a multi-gene editing recommendation tool is required. To achieve this goal, the optimization rules need to be set first. Then, on the basis of those existing single gene editing methods, the corresponding multi-gene editing tools could

be designed.

- **Other types of gene editing systems**

  CRISPR/Cas9 is one of the most widely applied gene editing systems. We understand it better comparing to the other CRISPR-based systems. However, studying other types of gene editing systems is also important. Firstly, different systems have their advantages. For example, the CRISPR/Cpf1 system contains smaller and simpler endonuclease (Cpf1), thus it has less limitations comparing to the Cas9 protein (Zetsche, Gootenberg, Abudayyeh, Slaymaker, Makarova, Essletzbichler, Volz, Joung, van der Oost, Regev et al. 2015, Tang, Lowder, Zhang, Malzahn, Zheng, Voytas, Zhong, Chen, Ren, Li et al. 2017). The CRISPR/Cas13 can edit RNA but not DNA, where we can edit the gene products without changing the genome (Cox, Gootenberg, Abudayyeh, Franklin, Kellner, Joung & Zhang 2017, Abudayyeh, Gootenberg, Essletzbichler, Han, Joung, Belanto, Verdine, Cox, Kellner, Regev et al. 2017). The base editor that generated by engineering fusions of CRISPR/Cas9 and a cytidine deaminase enzyme can correct single base without DNA strand break (Komor, Kim, Packer, Zuris & Liu 2016, Kim, Komor, Levy, Packer, Zhao & Liu 2017). These gene editing systems' designing tools are rare but necessary. Their efficiency evaluation and off-target effect problems cannot be ignored (Kim, Min, Song, Jung, Choi, Kim, Lee, Yoon & Kim 2018, Gehrke, Cervantes, Clement, Wu, Zeng, Bauer, Pinello & Joung 2018).

- **Disease oriented gene editing for precise gene therapy**

  A disease's development is a complex process and associates with a lot of genes. We hope to design a disease-oriented gene editing optimization system to help for precise gene therapy. It combines my background knowledge of disease-gene related research and gene-editing tool design experience. For a specific disease, its related gene regulation

network will be constructed first. Then, the network should be analyzed deeply to select one or more interested genes as the candidates. At last, our gene-editing design tools are used for the optimization and recommendation.

# Appendix A

# Appendix: Methodology foundation

## A.1 Adopted Mathematical and Statistical Conceptions

### A.1.1 Information entropy

In information theory, the information entropy is used to measure an event's uncertainty, where this event relates to a given probability distribution (Jaynes 1957, Cover & Thomas 1991). To characterize diseases for investigating the disease-lncRNA associations (Chapter 5), the information entropy was applied to represent the disease genes' distribution on the chromosome substructures and the disease enriched pathways' distribution on the manually generated pathway groups.

For a series $X$=$x_1$,$x_2$,$\cdots$,$x_i$,$\cdots$,$x_n$, its information entropy can be computed via below definition (formula A.1):

$$IE\left(X\right) = -\sum_{i=1}^{n} f_{x_i} log\left(f_{x_i}\right) \tag{A.1}$$

In formula A.1, $f_{x_i}$ means the frequency of $x_i$ in the series $X$. The base

of the logarithm is always set as 2 where the elements in $X$ is either 1 or 0.

### A.1.2 Fisher's exact test

Fisher's exact test was proposed by Ronald Fisher (Fisher 1922) to measure the significance of an event's happening when the observation number is small and the observations are presented with a 2*2 contingency table (Bower 2003). This work adopted the fisher's exact test to do disease gene pathway enrichment analysis, which helps the disease vectorization (Chapter 5).

The p-value of the fisher's exact test for indicating whether the statistic is significant can be computed via the hypergeometric distribution. For an observed 2*2 contingency table shown in Table A.1, the p-value can be obtained according to the following formula A.2:

Table A.1: **The example 2∗2 contingency table**

| | category A | category B |
|---|---|---|
| observation 1 | a | b |
| observation 1 | c | d |

$$[h, p, stats] = fishertest\left([a, b; c, d]\right) \tag{A.2}$$

In the formula A.2, the function 'fishertest' can be called with the Matlab software (version R2014b or higher). The output $p$ is the p-value. We computed them with default settings.

### A.1.3 Two-sample Kolmogorov−Smirnov test

As was defined by Lin et al. (Lin, Wu & Watada 2010), the Two-sample Kolmogorov-Smirnov test (two-sample K-S test, (Lilliefors 1967)) is a goodness-of-fit test to determine whether two underlying one-dimensional probability distributions differ. In this work, the two-sample K-S test was applied in two ways: to assess important features' distribution differences

(Chapter 6) and to help for feature space construction (Chapter 7). The output p-value of the two-sample K-S test shows the significance of the difference (p-value<0.05 means significant). It can be computed with the Matlab function 'kstest2' (version R2006a or higher).

## A.2 Applied Machine Learning Algorithms

### A.2.1 Support vector machine

For a dataset containing $n$ samples $x_i$, $i \in \{1, 2, \ldots, n\}$, where the label of $x_i$ is $y_i = 1$ if it is positive, otherwise $y_i = 0$, the SVM implemented by (Chang & Lin 2011) (Libsvm v3.22) can be expressed as the following dual formulation:

$$min_\alpha(L(\alpha)) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{j=1}^{n} \alpha_j$$

$$s.t. \sum_{i=1}^{n} y_i \alpha_i = 0, 0 \leqslant \alpha_i \leqslant C.$$

where $\alpha_i$ are Lagrange multipliers and $x_i$ for which $\alpha_i > 0$ are support vectors. The parameter $C$ controls the fraction of support vectors. $K(x_i, x_j)$ is the kernel function. The Radial Basis Function (RBF) kernel is applied to build the classifier, i.e.,

$$k_\sigma^{RBF} = exp(-\frac{1}{\sigma}||x_i - x_{j'}||^2)$$

where $\sigma$ is a parameter that controls the width of the radial basis. The parameters $C$ and $\sigma$ can be optimized with cross-validations (see next section).

### A.2.2 Ensemble SVM

Ensemble learning is one of the effective strategies to improve the machine learning performance and has been widely applied to various unbalanced

dataset classification problems (Yang et al. 2015). For $T$ base classifiers $f^i(x) = p(y = 1|x), i = 1, 2, \ldots, T$, where $f^i(x)$ approximates the probability of sample $x$ to be positive, the ensemble formulation is:

$$p^E(y = 1|x) = \frac{1}{T}\sum_{i=1}^{T} f^i(x)$$

For a given threshold $p_{threshold}$, if $p^E(y = 1|x) > p_{threshold}$ then, $x$ is labeled as 1, otherwise 0.

### A.2.3 XGBoost

XGBoost is a scalable end-to-end tree boosting system which was proposed by Chen et al. (Chen & Guestrin 2016). The authors have pointed out that their algorithm has two main novelties: it is a sparsity-aware algorithm; it applies weighted quantile sketch for approximate tree learning. In Chapter 6, we adopted XGBoost to conduct our first step regression of the CRISPR/Cas9 on-target cutting efficiencies. At the same time, the output feature importances guided the feature selection for our second step regression with SVM. We used the python package of the XGBoost downloaded from the website: https://github.com/dmlc/xgboost. More detail information about this algorithm can be found from their conference paper (Chen & Guestrin 2016).

## A.3 Cross-validation and Performance Indicators

### A.3.1 Cross-validation

Cross-validation is a widely used strategy for assessing how the constructed model can be generalized to an independent data set (see the introduction in Wikipedia, https://en.wikipedia.org/wiki/Cross-validation_(statistics)). It always splits the original dataset into 2 or more parts, where each part is adopted as a test set in turn to evaluate the model trained by the

remain data. This method can not only be used to evaluate the model but also to select optimal parameters for this model (Golub, Heath & Wahba 1979, Kohavi 1995).

The commonly applied cross-validations include n-fold cross-validation (n-fold cv, n=3, 5 or 10, see Chapters 3&5&6), leave-one-out cross-validation (loocv, see Chapters 3&5 ) and leave-one-group-out cross-validation (logocv, see Chapters 5&6&7). The only difference is the partition process. For example, in the n-fold cv, the original dataset is split into $n$ parts. In loocv, each sample is applied as a test sample in each round. However, in logocv, the partition is related to the group labels. For example, in Chapter 6, the leave-one-gene-out cv was adopted. The original dataset was split by the samples related genes (grouped by gene name). If the samples belong to total $N$ genes, the dataset is split into $N$ parts.

### A.3.2 Performance indicators

The machine learning model's performance can be evaluated by various indicators. Usually, the indicators for a classification problem are different from the regression issue.

For a classification question, the popular indicators include Accuracy, Precision, Specificity, Recall, F1 score, Matthews correlation coefficient (MCC), area under the ROC curve (AUROC) and area under the PR curve (AUPRC). The AUROC and AUPRC can be computed via the ROC curve and the PR curve. The Matlab function 'perfcurve' can be called to generate the ROC curve or PR curve and output the corresponding AUROC and AUPRC value. The python package 'scikit-learn' can also be used instead. The other 6 indicators can be computed via the following formulas:

$$Specificity = \frac{TN}{TN + FP} \tag{A.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{A.4}$$

$$Precision = \frac{TP}{TP + FP} \tag{A.5}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{A.6}$$

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{A.7}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{A.8}$$

where $TP, TN, FP$ and $FN$ represent true positive, true negative, false positive and false negative respectively. For these classifier indicators, the bigger the value, the better the performance.

The regression model's performance can be evaluated with the two widely used indicators such as the Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (SCC). For two given variables $X = \{x_1, x_2, \cdots, x_n\}$ and $Y = \{y_1, y_2, \cdots, y_n\}$, PCC and SCC can be calculated as following formulas:

$$PCC(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{A.9}$$

$$SCC(X, Y) = PCC(r_X, r_Y) \tag{A.10}$$

where, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. $r_X$ and $r_Y$ are the ranks of $X$ an $Y$ respectively.

Both of PCC and SCC have the value between -1 and 1, where 1 means the strongest positive correlation while -1 represents the strongest negative correlation.

# Appendix B

# Appendix: The list of databases that has been visited about this work

| name | description | authors (with website link) |
|---|---|---|
| HMDD v2.0 | human miRNA-disease associations | Li et al. (2013) |
| miR2Disease | human miRNA-disease associations | Jiang et al. (2009) |
| LncRNADisease | lncRNA-disease association | Chen et al. (2013) |
| Lnc2Cancer | Human LncRNA and Cancer Associations | Ning et al. (2016) |
| DO | human disease ontology | Schriml et al. (2012) |
| miRBase | miRNA database | Griffiths-Jones et al. (2004) |
| MeSH | Medical Subject Headings | Lipscomb et al. (2000) |
| CTD | Comparative Toxicogenomics Database | Mattingly et al. (2003) |
| HGNC | human genome resource | Povey et al. (2001) |
| SIDD | disease resources | Cheng et al. (2013) |
| miRecords | miRNA-target interaction resource | Xiao et al. (2009) |
| miRTarBase | miRNA-target interaction resource | Hsu et al. (2011) |
| miRCancer | miRNA-cancer associations | Xie et al. (2013) |
| GEO | gene expression profiles | Edgar et al. (2002) |
| LIBSVM | svm library | Chang et al. (2011) |
| OMIM | Human Genes and Genetic Disorders | Hamosh et al. (2000) |
| KEGG | pathway resources | Kanehisa et al. (2000) |
| DAVID | functional interpretation of large lists of genes | Dennis et al. (2003) |
| MNDR | ncRNA-disease associations in mammals | Wang et al. (2013) |
| RefSeq | NCBI Reference Sequence Database | Pruitt et al. (2000) |
| NONCODE | non-coding RNA resource | Liu et al. (2005) |
| Lncipedia | lncRNA sequence and annotation | Volders et al. (2012) |
| ensembl | genome browser for vertebrate genomes | Hubbard et al. (2002) |
| lncRNAdb | Functional lncRNA reference resource | Amaral et al. (2010) |
| Expression Atlas | gene and protein expression profiles | Kapushesky et al. (2011) |
| DisgeNet | disease gene resources | Bauer-Mehren et al. (2010) |
| malaCard | human maladies and their annotations | Rappaport et al. (2013) |

# Appendix C

# Appendix: List of Supplementary files

The Supplementary file list and the corresponding download links (https://drive.google.com/drive/folders/1YmayRRWw-9e0TmJg56OgteA-vRbYfQ3s?usp=sharing)

| name | chapter | description | link |
|---|---|---|---|
| Supplementary file 1 | 3 | disease genes | download |
| Supplementary file 2 | 3,4 | miRNA targets | download |
| Supplementary file 3 | 3 | GEO accessions | download |
| Supplementary file 4 | 3 | disease-miRNA association datasets | download |
| Supplementary file 5 | 3,4 | supplementary contents for chapter 3 and 4 | download |
| Supplementary file 6 | 3 | supplementary codes for chapter 3 and 4 | download |
| Supplementary file 7 | 3 | disease-miRNA positive samples | download |
| Supplementary file 8 | 4 | DGRs and predicted disease-miRNAs | download |
| Supplementary file 9 | 5 | disease genes | download |
| Supplementary file 10 | 5 | gene and pathway information | download |
| Supplementary file 11 | 5 | disease-lncRNAs and gene expression profiles | download |
| Supplementary file 12 | 5 | supplementary codes for chapter 4 | download |
| Supplementary file 13 | 5 | supplementary contents for chapter 4 | download |
| Supplementary file 14 | 5 | disease-similarity datasets | download |
| Supplementary file 15 | 6 | supplementary contents for chapter 6 | download |
| Supplementary file 16 | 6 | on-target site cutting efficiency datasets | download |
| Supplementary file 17 | 6 | U6 and T7 expression system test datasets | download |
| Supplementary file 18 | 7 | off-target site detection datasets | download |
| Supplementary file 19 | 7 | supplementary contents for chapter 7 | download |

# Appendix D

# Appendix: List of Symbols

The following list is neither exhaustive nor exclusive, but may be helpful.

| | |
|---|---|
| $CRISPR/Cas9$ | Clustered regularly interspaced short palindromic repeats/ CRISPR-associated protein 9 |
| $DNA$ | Deoxyribonucleic acid |
| $RNA$ | Ribonucleic acid |
| $miRNA$ | microRNA |
| $lncRNA$ | long non-coding RNA |
| $SVM$ | support vector machine |
| $DGR$ | disease-gene-microRNA |
| $mRNA$ | messenger RNA |
| $rRNA$ | ribosomal RNA |
| $snoRNA$ | small nucleolar RNA |
| $lincRNA$ | large intergenic (or intervening) noncoding RNA |
| $ncRNA$ | non-coding RNA |

| | |
|---|---|
| $BACE1$ | beta-site amyloid precursor protein cleaving enzyme 1 |
| $p53$ | Tumor protein p53 |
| $HMDD$ | the Human microRNA Disease Database |
| $ZFN$ | zinc finger nucleases |
| $TALEN$ | transcription activator-like effector nucleases |
| $sgRNA$ | single-guide RNA |
| $crRNA$ | CRISPR-RNA |
| $tracrRNA$ | trans-activation RNA |
| $PAM$ | protospacer adjacent motif |
| $HDR$ | homology-directed repair |
| $NHEJ$ | non-homology end joining |
| $DSB$ | double-strand break |
| $BCL2$ | B-cell lymphoma 2 |
| $RWR$ | random walk with restart |
| $PPI$ | protein-protein interaction |
| $NMF$ | non-negative matrix factorization |
| $PCR$ | Polymerase chain reaction |
| $G$ | guanine |
| $C$ | cytosine |
| $Cas9_{Sp}$ | Cas9 protein from Streptococcus pyogenes |
| $Cas9_{St1}$ | Cas9 protein from Streptococcus thermophilus |

| | |
|---|---|
| $PR$ | Precision-Recall |
| $dsODN$ | double-stranded oligodeoxynucleotide |
| $HTGTS$ | high-throughput, genome-wide translocation sequencing |
| $LAM-PCR$ | linear-amplificationmediated PCR |
| $NGS$ | next-generation sequencing |
| $DCDNN$ | deep convolutionary denosing neural network |
| $CNN$ | convolutionary neural network |
| $ROC$ | Receiver operating characteristic |
| $AUC$ | area under the ROC curve |
| $DO$ | Disease Ontology |
| $MeSH$ | Medical Subject Headings |
| $CTD$ | Comparative Toxicogenomics Database |
| $HGNC$ | The HUGO gene nomenclature committee (HGNC) |
| $SIDD$ | semantically integrated database |
| $GEO$ | Gene Expression Omnibus |
| $GSE$ | GEO accession |
| $Libsvm$ | A Library for Support Vector Machines |
| $DisSim$ | disease similarity |
| $MiRSim$ | miRNA similarity |
| $AvgDisSim$ | average disease similarity |
| $AvgMiRSim$ | average miRNA similarity |

| | |
|---|---|
| $TKM$ | training kernel matrix |
| $PKM$ | testing kernel matrix |
| $SemSim$ | disease semantic similarity |
| $FunSim$ | disease related genes' functional similarity |
| $DOSE$ | Disease Ontology Semantic and Enrichment analysis |
| $SeqSim$ | miRNA sequence similarity |
| $funSim$ | miRNA function similarity |
| $KMT$ | kernel matrix type |
| $LOOCV$ | leave-one-out cross-validation |
| $MTSS1$ | Metastasis suppressor protein 1 |
| $LOXL2$ | Lysyl oxidase homolog 2 |
| $qRT-PCR$ | quantitative reverse transcription polymerase chain reaction |
| $CLL$ | chronic lymphocytic leukaemia |
| $PTEN$ | Phosphatase and tensin homolog |
| $SMAD7$ | Mothers against decapentaplegic homolog 7 |
| $OMIM$ | Online Mendelian Inheritance In Man |
| $cfscore$ | multi-disease associated miRNA pair co-function score |
| $CDKN1A$ | cyclin-dependent kinase inhibitor 1 |
| $CCND1$ | Cyclin D1 |
| $KEGG$ | Kyoto Encyclopedia of Genes and Genomes |

| | |
|---|---|
| $DAVID$ | database for annotation, visualization and integrated discovery |
| $GO$ | gene ontology |
| $VEGFA$ | Vascular endothelial growth factor A |
| $MTHFR$ | Methylene tetrahydrofolate reductase |
| $IFNG$ | Interferon gamma |
| $RARB$ | Retinoic acid receptor beta |
| $MMP2$ | matrix metalloproteinase-2 |
| $MDM2$ | Mouse double minute 2 homolog |
| $CASP8$ | Caspase-8 |
| $HDAC3$ | Histone deacetylase 3 |
| $EZH2$ | Enhancer of zeste homolog 2 |
| $PU-learning$ | positive-unlabeled learning |
| $IBD$ | Inflammatory bowel disease |
| $LODOCV$ | leave-one-disease-out cross-validation |
| $RBF$ | Radial basis function |
| $SVM$ | support vector machine |
| $UCA1$ | Urothelial cancer associated 1 |
| $DLEU2$ | Deleted in lymphocytic leukemia 1 |
| $HOTAIR$ | HOXtranscript antisense RNA |
| $TSAM$ | two-step averaging method |

| | |
|---|---|
| $pHMM$ | profile Hidden Markov Model |
| $RS2$ | Rule set 2 |
| $FC$ | flow cytometry |
| $RES$ | drug resistance detection |
| $spCas9$ | Cas9 protein from Streptococcus pyogenes |
| $stlCas9$ | Cas9 protein from Streptococcus thermophilus |
| $UTR$ | untranslated region |
| $MT$ | Mutation Type |
| $MCC$ | Matthews correlation coefficient |
| $LOGOCV$ | leave-one-gene-out cross-validation |
| $CYBB$ | NADPH oxidase 2 |
| $onTSeq$ | on-target site sequence |
| $offTSeq$ | off-target site sequence |
| $noEdSeq$ | no-editing target site sequence |
| $canSeq$ | candidate target site sequence |
| $Two-sample\ K-S\ test$ | Two-sample Kolmogorov-Smirnov test |
| $NCC$ | nucleotide composition change features |
| $logocv$ | leave-one-guide-out cross-validation |
| $AUROC$ | area under the ROC curve |
| $AUPRC$ | area under the PR curve |
| $EMX1$ | empty spiracles homeobox 1 |

| | |
|---|---|
| $OR$ | overlap rate |
| $SNP$ | single nucleotide polymorphism |
| $Cpf1$ | simpler endonuclease |

# Bibliography

Abadi, S., Yan, W. X., Amar, D. & Mayrose, I. (2017), 'A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action', *PLoS Computational Biology* **13**(10), e1005807.

Abudayyeh, O. O., Gootenberg, J. S., Essletzbichler, P., Han, S., Joung, J., Belanto, J. J., Verdine, V., Cox, D. B., Kellner, M. J., Regev, A. et al. (2017), 'RNA targeting with CRISPR–Cas13', *Nature* **550**(7675), 280–284.

Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., Garca Girn, C., Hourlier, T., Howe, K., Khri, A., Kokocinski, F., Martin, F. J., Murphy, D. N., Nag, R., Ruffier, M., Schuster, M., Tang, Y. A., Vogel, J.-H., White, S., Zadissa, A., Flicek, P. & Searle, S. M. J. (2016), 'The Ensembl gene annotation system', *Database* **2016**, baw093.
**URL:** *http://dx.doi.org/10.1093/database/baw093*

Alaimo, S., Giugno, R. & Pulvirenti, A. (2014), 'ncPred: ncRNA-disease association prediction through tripartite network-based inference', *Frontiers in Bioengineering and Biotechnology* **2**, 71.

Alberts, B., Johnson, A., Lewis, J., Walter, P., Raff, M. & Roberts, K. (2002), 'Molecular Biology of the Cell 4th Edition: International Student Edition'.

Ambros, V., Bartel, B., Bartel, D. P., Burge, C. B., Carrington, J. C., Chen, X., Dreyfuss, G., Eddy, S. R., Griffiths-Jones, S., Marshall, M. et al. (2003), 'A uniform system for microRNA annotation', *RNA* **9**(3), 277–279.

Amodio, N., Rossi, M., Raimondi, L., Pitari, M. R., Botta, C., Tagliaferri, P. & Tassone, P. (2015), 'miR-29s: a family of epi-miRNAs with therapeutic implications in hematologic malignancies', *Oncotarget* **6**(15), 12837.

Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. (2014), 'Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease', *Nature* **513**(7519), 569–573.

Bae, S., Park, J. & Kim, J.-S. (2014), 'Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases', *Bioinformatics* **30**(10), 1473–1475.

Bartel, D. P. (2009), 'MicroRNAs: target recognition and regulatory functions', *Cell* **136**(2), 215–233.

Bartel, D. P. (2018), 'Metazoan MicroRNAs', *Cell* **173**(1), 20–51.

Barutcuoglu, Z., Schapire, R. E. & Troyanskaya, O. G. (2006), 'Hierarchical multi-label prediction of gene function', *Bioinformatics* **22**(7), 830–836.

Baskerville, S. & Bartel, D. P. (2005), 'Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes', *RNA* **11**(3), 241–247.

Bauer-Mehren, A., Rautschka, M., Sanz, F. & Furlong, L. I. (2010), 'DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene–disease networks', *Bioinformatics* **26**(22), 2924–2926.

Beißbarth, T. & Speed, T. P. (2004), 'GOstat: find statistically overrepresented Gene Ontologies within a group of genes', *Bioinformatics* **20**(9), 1464–1465.

Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R. H. & Cuppen, E. (2005), 'Phylogenetic shadowing and computational identification of human microRNA genes', *Cell* **120**(1), 21–24.

Bindewald, E. & Shapiro, B. A. (2006), 'RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers', *RNA* **12**(3), 342–352.

Biswas, A. K., Gao, J. X., Zhang, B. & Wu, X. (2014), NMF-Based LncRNA-Disease Association Inference and Bi-Clustering, *in* '2014 IEEE International Conference on Bioinformatics and Bioengineering', pp. 97–104.

Bock, J. R. & Gough, D. A. (2001), 'Predicting protein–protein interactions from primary structure', *Bioinformatics* **17**(5), 455–460.

Bower, K. M. (2003), When to use Fisher's exact test, *in* 'American Society for Quality, Six Sigma Forum Magazine', Vol. 2, pp. 35–37.

Broderick, J. A., Salomon, W. E., Ryder, S. P., Aronin, N. & Zamore, P. D. (2011), 'Argonaute protein identity and pairing geometry determine cooperativity in mammalian RNA silencing', *RNA* **17**(10), 1858–1869.

Cai, C., Han, L., Ji, Z. L., Chen, X. & Chen, Y. Z. (2003), 'SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence', *Nucleic Acids Research* **31**(13), 3692–3697.

Cameron, P., Fuller, C. K., Donohoue, P. D., Jones, B. N., Thompson, M. S., Carter, M. M., Gradia, S., Vidal, B., Garner, E., Slorach, E. M. et al. (2017), 'Mapping the genomic landscape of CRISPR–Cas9 cleavage', *Nature Methods* **14**(6), 600–606.

Carpenter, S., Aiello, D., Atianand, M. K., Ricci, E. P., Gandhi, P., Hall, L. L., Byron, M., Monks, B., Henry-Bezy, M., Lawrence, J. B. et al. (2013), 'A long noncoding RNA mediates both activation and repression of immune response genes', *Science* **341**(6147), 789–792.

Chang, C.-C. & Lin, C.-J. (2011), 'LIBSVM: a library for support vector machines', *ACM Transactions on Intelligent Systems and Technology (TIST)* **2**(3), 27.

Chapelle, O., Scholkopf, B. & Zien, A. (2009), 'Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]', *IEEE Transactions on Neural Networks* **20**(3), 542–542.

Chari, R., Mali, P., Moosburner, M. & Church, G. M. (2015), 'Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach', *Nature Methods* **12**(9), 823–826.

Chari, R., Yeo, N. C., Chavez, A. & Church, G. M. (2017), 'sgRNA Scorer 2.0: a species-independent model to predict CRISPR/Cas9 activity', *ACS Synthetic Biology* **6**(5), 902–904.

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., Zhang, Q., Yan, G. & Cui, Q. (2013), 'LncRNADisease: a database for long-non-coding RNA-associated diseases', *Nucleic Acids Research* **41**(D1), D983–D986.

Chen, H. & Zhang, Z. (2013), 'Similarity-based methods for potential human microRNA-disease association prediction', *BMC Medical Genomics* **6**(1), 12.

Chen, T. & Guestrin, C. (2016), XGBoost: A Scalable Tree Boosting System, *in* 'Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '16, ACM, New York, NY, USA, pp. 785–794.
**URL:** *http://doi.acm.org/10.1145/2939672.2939785*

Chen, X. (2015), 'KATZLDA: KATZ measure for the lncRNA-disease association prediction', *Scientific Reports* **5**, 16840.

Chen, X., Liu, M.-X. & Yan, G.-Y. (2012), 'RWRMDA: predicting novel human microRNA–disease associations', *Molecular BioSystems* **8**(10), 2792–2798.

Chen, X. & Yan, G.-Y. (2013), 'Novel human lncRNA–disease association inference based on lncRNA expression profiles', *Bioinformatics* **29**(20), 2617–2624.

Chen, X. & Yan, G.-Y. (2014), 'Semi-supervised learning for potential human microRNA-disease associations inference', *Scientific Reports* **4**, 5501.

Chen, X., Yang, C., Xie, S. & Cheung, E. (2018), 'Long non-coding RNA GAS5 and ZFAS1 are prognostic markers involved in translation targeted by miR-940 in prostate cancer', *Oncotarget* **9**(1), 1048.

Cheng, L., Li, J., Ju, P., Peng, J. & Wang, Y. (2014), 'SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association', *PLoS ONE* **9**(6), e99415.

Cheng, L., Wang, G., Li, J., Zhang, T., Xu, P. & Wang, Y. (2013), 'SIDD: a semantically integrated database towards a global view of human disease', *PLoS ONE* **8**(10), e75504.

Chiarle, R., Zhang, Y., Frock, R. L., Lewis, S. M., Molinie, B., Ho, Y.-J., Myers, D. R., Choi, V. W., Compagno, M., Malkin, D. J. et al. (2011), 'Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells', *Cell* **147**(1), 107–119.

Cho, S. W., Kim, S., Kim, Y., Kweon, J., Kim, H. S., Bae, S. & Kim, J.-S. (2014), 'Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases', *Genome Research* **24**(1), 132–141.

Chou, K.-C. (2001), 'Prediction of protein cellular attributes using pseudo-amino acid composition', *Proteins: Structure, Function, and Bioinformatics* **43**(3), 246–255.

Chuai, G., Ma, H., Yan, J., Chen, M., Hong, N., Xue, D., Zhou, C., Zhu, C., Chen, K., Duan, B. et al. (2018), 'DeepCRISPR: optimized CRISPR guide RNA design by deep learning', *Genome Biology* **19**(1), 80.

Cittelly, D. M., Das, P. M., Salvo, V. A., Fonseca, J. P., Burow, M. E. & Jones, F. E. (2010), 'Oncogenic HER2Δ16 suppresses miR-15a/16 and deregulates BCL-2 to promote endocrine resistance of breast tumors', *Carcinogenesis* **31**(12), 2049–2057.

Cittelly, D. M., Finlay-Schultz, J., Howe, E. N., Spoelstra, N. S., Axlund, S. D., Hendricks, P., Jacobsen, B. M., Sartorius, C. A. & Richer, J. K. (2013), 'Progestin suppression of miR-29 potentiates dedifferentiation of breast cancer cells via KLF4', *Oncogene* **32**(20), 2555–2564.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. et al. (2009), 'Biopython: freely available Python tools for computational molecular biology and bioinformatics', *Bioinformatics* **25**(11), 1422–1423.

Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Machine Learning* **20**(3), 273–297.

Cover, T. M. & Thomas, J. A. (1991), *Elements of Information Theory*, Wiley-Interscience, New York, NY, USA.

Cox, D. B., Gootenberg, J. S., Abudayyeh, O. O., Franklin, B., Kellner, M. J., Joung, J. & Zhang, F. (2017), 'RNA editing with CRISPR-Cas13', *Science* **358**(6366), 1019–1027.

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S. et al. (2014), 'Ensembl 2015', *Nucleic Acids Research* **43**(D1), D662–D669.

Davis, A. P., Murphy, C. G., Johnson, R., Lay, J. M., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B. L., Rosenstein, M. C., Wiegers, T. C. et al. (2013), 'The comparative toxicogenomics database: update 2013', *Nucleic Acids Research* **41**(D1), D1104–D1114.

Davis, A. P., Murphy, C. G., Saraceni-Richards, C. A., Rosenstein, M. C., Wiegers, T. C. & Mattingly, C. J. (2009), 'Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks', *Nucleic Acids Research* **37**(suppl 1), D786–D792.

Davis, J. & Goadrich, M. (2006), The Relationship Between Precision-Recall and ROC Curves, *in* 'Proceedings of the 23rd International Conference on Machine Learning', ICML '06, ACM, New York, NY, USA, pp. 233–240.
**URL:** *http://doi.acm.org/10.1145/1143844.1143874*

De Ravin, S. S., Li, L., Wu, X., Choi, U., Allen, C., Koontz, S., Lee, J., Theobald-Whiting, N., Chu, J., Garofalo, M. et al. (2017), 'CRISPR-Cas9 gene repair of hematopoietic stem cells from patients with X-linked chronic granulomatous disease', *Science Translational Medicine* **9**(372), eaah3480.

Delgado, A., Brandao, P. & Narayanan, R. (2014), 'Diabetes associated genes from the dark matter of the human proteome', *MOJ Proteomics Bioinform* **1**(4), 00020.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G. et al. (2012), 'The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression', *Genome Research* **22**(9), 1775–1789.

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R. et al.

(2016), 'Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9', *Nature Biotechnology* **34**(2), 184–191.

Doench, J. G., Hartenian, E., Graham, D. B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B. L., Xavier, R. J. & Root, D. E. (2014), 'Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation', *Nature Biotechnology* **32**(12), 1262–1267.

Doudna, J. A. & Charpentier, E. (2014), 'The new frontier of genome engineering with CRISPR-Cas9', *Science* **346**(6213), 1258096.

Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. (1998), *Biological sequence analysis: probabilistic models of proteins and nucleic acids*, Cambridge university press.

Dykes, I. M. & Emanueli, C. (2017), 'Transcriptional and post-transcriptional gene regulation by long non-coding RNA', *Genomics, Proteomics & Bioinformatics* **15**(3), 177–186.

Eddy, S. R. (1998), 'Profile hidden Markov models', *Bioinformatics* **14**(9), 755–763.

Eddy, S. R. (2001), 'Non–coding RNA genes and the modern RNA world', *Nature Reviews Genetics* **2**(12), 919–929.

Edgar, R., Domrachev, M. & Lash, A. E. (2002), 'Gene Expression Omnibus: NCBI gene expression and hybridization array data repository', *Nucleic Acids Research* **30**(1), 207–210.

Elkan, C. & Noto, K. (2008), Learning Classifiers from Only Positive and Unlabeled Data, *in* 'Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '08, ACM, New York, NY, USA, pp. 213–220.
**URL:** *http://doi.acm.org/10.1145/1401890.1401920*

Eulalio, A., Huntzinger, E., Nishihara, T., Rehwinkel, J., Fauser, M. & Izaurralde, E. (2009), 'Deadenylation is a widespread effect of miRNA regulation', *RNA* **15**(1), 21–32.

Fang, Y. & Fullwood, M. J. (2016), 'Roles, functions, and mechanisms of long non-coding RNAs in cancer', *Genomics, Proteomics & Bioinformatics* **14**(1), 42–54.

Farasat, I. & Salis, H. M. (2016), 'A biophysical model of CRISPR/Cas9 activity for rational design of genome editing and gene regulation', *PLoS Computational Biology* **12**(1), e1004724.

Fisher, R. A. (1922), 'On the interpretation of $\chi$ 2 from contingency tables, and the calculation of P', *Journal of the Royal Statistical Society* **85**(1), 87–94.

Forney, G. D. (1973), 'The viterbi algorithm', *Proceedings of the IEEE* **61**(3), 268–278.

Freudenberg, J. & Propping, P. (2002), 'A similarity-based method for genome-wide prediction of disease-relevant human genes', *Bioinformatics* **18**(suppl_2), S110–S115.

Frock, R. L., Hu, J., Meyers, R. M., Ho, Y.-J., Kii, E. & Alt, F. W. (2015), 'Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases', *Nature Biotechnology* **33**(2), 179.

Fu, B. X., St Onge, R. P., Fire, A. Z. & Smith, J. D. (2016), 'Distinct patterns of Cas9 mismatch tolerance in vitro and in vivo', *Nucleic Acids Research* **44**(11), 5365–5377.

Fu, Y., Foden, J. A., Khayter, C., Maeder, M. L., Reyon, D., Joung, J. K. & Sander, J. D. (2013), 'High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells', *Nature Biotechnology* **31**(9), 822–826.

Fusi, N., Smith, I., Doench, J. & Listgarten, J. (2015), 'In Silico Predictive Modeling of CRISPR/Cas9 guide efficiency', *BioRxiv* p. 021568.

Ganegoda, G. U., Li, M., Wang, W. & Feng, Q. (2015), 'Heterogeneous network model to infer human disease-long intergenic non-coding RNA associations', *IEEE Transactions on Nanobioscience* **14**(2), 175–183.

Gao, X., Tao, Y., Lamas, V., Huang, M., Yeh, W.-H., Pan, B., Hu, Y.-J., Hu, J. H., Thompson, D. B., Shu, Y. et al. (2018), 'Treatment of autosomal dominant hearing loss by in vivo delivery of genome editing agents', *Nature* **553**(7687), 217–221.

Gebert, L. F. & MacRae, I. J. (2018), 'Regulation of microRNA function in animals', *Nature Reviews Molecular Cell Biology* .

Gehrke, J. M., Cervantes, O., Clement, M. K., Wu, Y., Zeng, J., Bauer, D. E., Pinello, L. & Joung, J. K. (2018), 'An APOBEC3A-Cas9 base editor with minimized bystander and off-target activities', *Nature Biotechnology* **36**, 977982.

Ginno, P. A., Lim, Y. W., Lott, P. L., Korf, I. & Chédin, F. (2013), 'GC skew at the 5' and 3' ends of human genes links R-loop formation to epigenetic regulation and transcription termination', *Genome Research* **23**(10), 1590–1600.

Golub, G. H., Heath, M. & Wahba, G. (1979), 'Generalized cross-validation as a method for choosing a good ridge parameter', *Technometrics* **21**(2), 215–223.

Guo, J., Wang, T., Guan, C., Liu, B., Luo, C., Xie, Z., Zhang, C. & Xing, X.-H. (2018), 'Improved sgRNA design in bacteria via genome-wide activity profiling', *Nucleic Acids Research* **46**(14), 7052–7069.
**URL:** *http://dx.doi.org/10.1093/nar/gky572*

Guyon, I. & Elisseeff, A. (2003), 'An introduction to variable and feature selection', *Journal of Machine Learning Research* **3**(Mar), 1157–1182.

Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J.-B., Schneider-Maunoury, S., Shkumatava, A., Teboul, L., Kent, J. et al. (2016), 'Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR', *Genome Biology* **17**(1), 148.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. (2005), 'Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders', *Nucleic Acids Research* **33**(suppl_1), D514–D517.

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K. & Kjems, J. (2013), 'Natural RNA circles function as efficient microRNA sponges', *Nature* **495**(7441), 384.

Hao, Y., Colak, R., Teyra, J., Corbi-Verge, C., Ignatchenko, A., Hahne, H., Wilhelm, M., Kuster, B., Braun, P., Kaida, D. et al. (2015), 'Semi-supervised learning predicts approximately one third of the alternative splicing isoforms as functional proteins', *Cell Reports* **12**(2), 183–189.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), Unsupervised learning, *in* 'The Elements of Statistical Learning', Springer, pp. 485–585.

Hausser, J. & Zavolan, M. (2014), 'Identification and consequences of miRNA-target interactions [mdash] beyond repression of gene expression', *Nature Reviews Genetics* **15**(9), 599–612.

He, H. & Garcia, E. A. (2009), 'Learning from imbalanced data', *IEEE Transactions on knowledge and data engineering* **21**(9), 1263–1284.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. & Scholkopf, B. (1998), 'Support vector machines', *IEEE Intelligent Systems and Their Applications* **13**(4), 18–28.

Hébert, S. S., Horré, K., Nicolaï, L., Papadopoulou, A. S., Mandemakers, W., Silahtaroglu, A. N., Kauppinen, S., Delacourte, A. & De Strooper, B.

(2008), 'Loss of microRNA cluster miR-29a/b-1 in sporadic Alzheimer's disease correlates with increased BACE1/$\beta$-secretase expression', *Proceedings of the National Academy of Sciences* **105**(17), 6415–6420.

Ho, T.-T., Zhou, N., Huang, J., Koirala, P., Xu, M., Fung, R., Wu, F. & Mo, Y.-Y. (2014), 'Targeting non-coding RNAs with the CRISPR/Cas9 system in human cell lines', *Nucleic Acids Research* **43**(3), e17–e17.

Hoehndorf, R., Schofield, P. N. & Gkoutos, G. V. (2015), 'Analysis of the human diseasome using phenotype similarity between common, genetic, and infectious diseases', *Scientific Reports* **5**, 10888.

Hsu, P. D., Lander, E. S. & Zhang, F. (2014), 'Development and applications of CRISPR-Cas9 for genome engineering', *Cell* **157**(6), 1262–1278.

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., Li, Y., Fine, E. J., Wu, X., Shalem, O. et al. (2013), 'DNA targeting specificity of RNA-guided Cas9 nucleases', *Nature Biotechnology* **31**(9), 827.

Hsu, S.-D., Lin, F.-M., Wu, W.-Y., Liang, C., Huang, W.-C., Chan, W.-L., Tsai, W.-T., Chen, G.-Z., Lee, C.-J., Chiu, C.-M. et al. (2010), 'miRTarBase: a database curates experimentally validated microRNA–target interactions', *Nucleic Acids Research* **39**(suppl_1), D163–D169.

Huang, D. W., Sherman, B. T. & Lempicki, R. A. (2009*a*), 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic Acids Research* **37**(1), 1–13.

Huang, D. W., Sherman, B. T. & Lempicki, R. A. (2009*b*), 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources', *Nature Protocols* **4**(1), 44–57.

Huang, Y.-A., Chen, X., You, Z.-H., Huang, D.-S. & Chan, K. C. (2016), 'ILNCSIM: improved lncRNA functional similarity calculation model', *Oncotarget* **7**(18), 25902.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T. et al. (2002), 'The Ensembl genome database project', *Nucleic Acids Research* **30**(1), 38–41.

Huo, L., Zhang, H., Huo, X., Yang, Y., Li, X. & Yin, Y. (2017), 'pHMM-tree: phylogeny of profile hidden Markov models', *Bioinformatics* **33**(7), 1093–1095.

Jaynes, E. T. (1957), 'Information theory and statistical mechanics', *Physical Review* **106**(4), 620.

Ji, X., Lu, H., Zhou, Q. & Luo, K. (2014), 'LARP7 suppresses P-TEFb activity to inhibit breast cancer progression and metastasis', *Elife* **3**, e02907.

Jiang, F. & Doudna, J. A. (2017), 'CRISPR–Cas9 structures and mechanisms', *Annual Review of Biophysics* **46**, 505–529.

Jiang, F., Taylor, D. W., Chen, J. S., Kornfeld, J. E., Zhou, K., Thompson, A. J., Nogales, E. & Doudna, J. A. (2016), 'Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage', *Science* **351**(6275), 867–871.

Jiang, H., Zhang, G., Wu, J.-H. & Jiang, C.-P. (2014), 'Diverse roles of miR-29 in cancer (review)', *Oncology Reports* **31**(4), 1509–1516.

Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., Teng, M., Liu, Y. & Wang, Y. (2010), 'Prioritization of disease microRNAs through a human phenome-microRNAome network', *BMC Systems Biology* **4**(1), S2.

Jiang, Q., Wang, G., Jin, S., Li, Y. & Wang, Y. (2013), 'Predicting human microRNA-disease associations based on support vector machine', *International Journal of Data Mining and Bioinformatics* **8**(3), 282–293.

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G. & Liu, Y. (2009), 'miR2Disease: a manually curated database for microRNA deregulation in human disease', *Nucleic Acids Research* **37**(suppl 1), D98–D104.

Jinek, M., Jiang, F., Taylor, D. W., Sternberg, S. H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S. et al. (2014), 'Structures of Cas9 endonucleases reveal RNA-mediated conformational activation', *Science* **343**(6176), 1247997.

Joung, J. K. & Sander, J. D. (2013), 'TALENs: a widely applicable technology for targeted genome editing', *Nature Reviews Molecular Cell Biology* **14**(1), 49.

Kan, Y., Ruis, B., Takasugi, T. & Hendrickson, E. A. (2017), 'Mechanisms of precise genome editing using oligonucleotide donors', *Genome Research* **27**(7), 1099–1111.

Kanchiswamy, C. N., Sargent, D. J., Velasco, R., Maffei, M. E. & Malnoy, M. (2015), 'Looking forward to genetically edited fruit crops', *Trends in Biotechnology* **33**(2), 62–64.

Kanehisa, M. & Goto, S. (2000), 'KEGG: kyoto encyclopedia of genes and genomes', *Nucleic Acids Research* **28**(1), 27–30.

Karplus, K., Barrett, C. & Hughey, R. (1998), 'Hidden Markov models for detecting remote protein homologies', *Bioinformatics* **14**(10), 846–856.

Kato, M., Kurozumi, A., Goto, Y., Matsushita, R., Okato, A., Nishikawa, R., Fukumoto, I., Koshizuka, K., Ichikawa, T. & Seki, N. (2017), 'Regulation of metastasis-promoting LOXL2 gene expression by antitumor microRNAs in prostate cancer', *Journal of Human Genetics* **62**(1), 123.

Kaur, K., Gupta, A. K., Rajput, A. & Kumar, M. (2016), 'ge-CRISPR-An integrated pipeline for the prediction and analysis of sgRNAs genome editing efficiency for CRISPR/Cas system', *Scientific Reports* **6**, 30870.

Kedmi, M., Ben-Chetrit, N., Körner, C., Mancini, M., Ben-Moshe, N. B., Lauriola, M., Lavi, S., Biagioni, F., Carvalho, S., Cohen-Dvashi, H. et al. (2015), 'EGF induces microRNAs that target suppressors of cell migration: miR-15b targets MTSS1 in breast cancer', *Science Signaling* **8**(368), ra29–ra29.

Kim, D., Bae, S., Park, J., Kim, E., Kim, S., Yu, H. R., Hwang, J., Kim, J.-I. & Kim, J.-S. (2015), 'Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells', *Nature Methods* **12**(3), 237.

Kim, D., Kim, S., Kim, S., Park, J. & Kim, J.-S. (2016), 'Genome-wide target specificities of CRISPR-Cas9 nucleases revealed by multiplex Digenome-seq', *Genome Research* **26**(3), 406–415.

Kim, H. K., Min, S., Song, M., Jung, S., Choi, J. W., Kim, Y., Lee, S., Yoon, S. & Kim, H. H. (2018), 'Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity', *Nature Biotechnology* **36**(3), 239.

Kim, V. N. (2005), 'Small RNAs: classification, biogenesis, and function', *Molecules and Cells* **19**(1), 1–15.

Kim, Y. B., Komor, A. C., Levy, J. M., Packer, M. S., Zhao, K. T. & Liu, D. R. (2017), 'Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions', *Nature Biotechnology* **35**(4), 371.

Kim, Y.-K. & Kim, V. N. (2007), 'Processing of intronic microRNAs', *The EMBO journal* **26**(3), 775–783.

Kleinstiver, B. P., Pattanayak, V., Prew, M. S., Tsai, S. Q., Nguyen, N. T., Zheng, Z. & Joung, J. K. (2016), 'High-fidelity CRISPR–Cas9

nucleases with no detectable genome-wide off-target effects', *Nature* **529**(7587), 490–495.

Kohavi, R. (1995), A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection, *in* 'Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2', IJCAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 1137–1143.
**URL:** *http://dl.acm.org/citation.cfm?id=1643031.1643047*

Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. (2016), 'Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage', *Nature* **533**(7603), 420.

Kotsiantis, S. B., Zaharakis, I. & Pintelas, P. (2007), 'Supervised machine learning: A review of classification techniques', *Emerging Artificial Intelligence Applications in Computer Engineering* **160**, 3–24.

Kozomara, A. & Griffiths-Jones, S. (2014), 'miRBase: annotating high confidence microRNAs using deep sequencing data', *Nucleic Acids Research* **42**(D1), D68–D73.

Kramer, N. J., Haney, M. S., Morgens, D. W., Jovičić, A., Couthouis, J., Li, A., Ousey, J., Ma, R., Bieri, G., Tsui, C. K. et al. (2018), 'CRISPR–Cas9 screens in human cells and primary neurons identify modifiers of C9ORF72 dipeptide-repeat-protein toxicity', *Nature Genetics* **50**(4), 603.

Kumarswamy, R., Bauters, C., Volkmann, I., Maury, F., Fetisch, J., Holzmann, A., Lemesle, G., de Groote, P., Pinet, F. & Thum, T. (2014), 'Circulating long noncoding RNA, LIPCAR, predicts survival in patients with heart failure', *Circulation Research* **114**(10), 1569–1575.

Kung, J. T., Colognori, D. & Lee, J. T. (2013), 'Long noncoding RNAs: past, present, and future', *Genetics* **193**(3), 651–669.

Lai, X., Schmitz, U., Gupta, S. K., Bhattacharya, A., Kunz, M., Wolkenhauer, O. & Vera, J. (2012), 'Computational analysis of target hub gene repression regulated by multiple and cooperative miRNAs', *Nucleic Acids Research* **40**(18), 8818–8834.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. (2009), 'Ultrafast and memory-efficient alignment of short DNA sequences to the human genome', *Genome Biology* **10**(3), R25.

Le, D.-H., Hoai, N. X. & Kwon, Y.-K. (2015), A Comparative study of classification-based machine learning methods for novel disease gene prediction, *in* 'Knowledge and Systems Engineering', Springer, pp. 577–588.

Le Novere, N. (2001), 'MELTING, computing the melting temperature of nucleic acid duplex', *Bioinformatics* **17**(12), 1226–1227.

LeCun, Y., Bengio, Y. & Hinton, G. (2015), 'Deep learning', *Nature* **521**(7553), 436–444.

Lee, C. M., Cradick, T. J. & Bao, G. (2016), 'The Neisseria meningitidis CRISPR-Cas9 system enables specific genome editing in mammalian cells', *Molecular Therapy* **24**(3), 645–654.

Lerner, M., Harada, M., Lovén, J., Castro, J., Davis, Z., Oscier, D., Henriksson, M., Sangfelt, O., Grandér, D. & Corcoran, M. M. (2009), 'DLEU2, frequently deleted in malignancy, functions as a critical host gene of the cell cycle inhibitory microRNAs miR-15a and miR-16-1', *Experimental Cell Research* **315**(17), 2941–2952.

Li, J., Gao, C., Wang, Y., Ma, W., Tu, J., Wang, J., Chen, Z., Kong, W. & Cui, Q. (2014), 'A bioinformatics method for predicting long noncoding RNAs associated with vascular disease', *Science China Life sciences* **57**(8), 852–857.

Li, J., Gong, B., Chen, X., Liu, T., Wu, C., Zhang, F., Li, C., Li, X., Rao, S. & Li, X. (2011), 'DOSim: an R package for similarity between diseases based on disease ontology', *BMC Bioinformatics* **12**(1), 266.

Li, X., Deng, S.-j., Zhu, S., Jin, Y., Cui, S.-p., Chen, J.-y., Xiang, C., Li, Q.-y., He, C., Zhao, S.-f. et al. (2016), 'Hypoxia-induced lncRNA-NUTF2P3-001 contributes to tumorigenesis of pancreatic cancer by derepressing the miR-3923/KRAS pathway', *Oncotarget* **7**(5), 6000.

Li, X. & Liu, B. (2003), Learning to Classify Texts Using Positive and Unlabeled Data, *in* 'Proceedings of the 18th International Joint Conference on Artificial Intelligence', IJCAI'03, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 587–592.
**URL:** *http://dl.acm.org/citation.cfm?id=1630659.1630746*

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T. & Cui, Q. (2013), 'HMDD v2. 0: a database for experimentally supported human microRNA and disease associations', *Nucleic Acids Research* **42**(D1), D1070–D1074.

Liaw, A., Wiener, M. et al. (2002), 'Classification and regression by randomForest', *R News* **2**(3), 18–22.

Lilliefors, H. W. (1967), 'On the Kolmogorov-Smirnov test for normality with mean and variance unknown', *Journal of the American Statistical Association* **62**(318), 399–402.

Lin, P.-C., Wu, B. & Watada, J. (2010), Kolmogorov-Smirnov two sample test with continuous fuzzy data, *in* 'Integrated Uncertainty Management and Applications', Springer, pp. 175–186.

Ling, H., Fabbri, M. & Calin, G. A. (2013), 'MicroRNAs and other non-coding RNAs as targets for anticancer drug development', *Nature Reviews Drug Discovery* **12**(11), 847–865.

Linghu, B., Snitkin, E. S., Hu, Z., Xia, Y. & DeLisi, C. (2009), 'Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network', *Genome Biology* **10**(9), 1–17.

Lipscomb, C. E. (2000), 'Medical subject headings (MeSH)', *Bulletin of the Medical Library Association* **88**(3), 265.

Listgarten, J., Weinstein, M., Kleinstiver, B. P., Sousa, A. A., Joung, J. K., Crawford, J., Gao, K., Hoang, L., Elibol, M., Doench, J. G. et al. (2018), 'Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs', *Nature Biomedical Engineering* **2**(1), 38.

Liu, B., Dai, Y., Li, X., Lee, W. S. & Yu, P. S. (2003), Building text classifiers using positive and unlabeled examples, *in* 'Third IEEE International Conference on Data Mining', pp. 179–186.

Liu, M.-X., Chen, X., Chen, G., Cui, Q.-H. & Yan, G.-Y. (2014), 'A computational framework to infer human disease-associated long noncoding RNAs', *PLoS ONE* **9**(1), e84408.

Liu, Q., Huang, J., Zhou, N., Zhang, Z., Zhang, A., Lu, Z., Wu, F. & Mo, Y.-Y. (2013), 'LncRNA loc285194 is a p53-regulated tumor suppressor', *Nucleic Acids Research* **41**(9), 4976–4987.

Liu, X., Li, D., Zhang, W., Guo, M. & Zhan, Q. (2012), 'Long non-coding RNA gadd7 interacts with TDP-43 and regulates Cdk6 mRNA decay', *The EMBO journal* **31**(23), 4415–4427.

Liu, Z.-P., Wu, L.-Y., Wang, Y., Zhang, X.-S. & Chen, L. (2010), 'Prediction of protein–RNA binding sites by a random forest method with combined features', *Bioinformatics* **26**(13), 1616–1622.

Liu, Z., Yang, D., Xie, P., Ren, G., Sun, G., Zeng, X. & Sun, X. (2012), 'MiR-106b and MiR-15b modulate apoptosis and angiogenesis in myocardial infarction', *Cellular Physiology and Biochemistry* **29**(5-6), 851–862.

Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W. & Cui, Q. (2008), 'An analysis of human microRNA and disease associations', *PLoS ONE* **3**(10), e3420.

MacFarlane, L.-A. & R Murphy, P. (2010), 'MicroRNA: biogenesis, function and role in cancer', *Current Genomics* **11**(7), 537–561.

Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. (2005), 'Entrez Gene: gene-centered information at NCBI', *Nucleic Acids Research* **33**(suppl_1), D54–D58.

Mao, K. Z. (2004), 'Orthogonal forward selection and backward elimination algorithms for feature subset selection', *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **34**(1), 629–634.

Marchese, F. P., Raimondi, I. & Huarte, M. (2017), 'The multidimensional mechanisms of long noncoding RNA function', *Genome Biology* **18**(1), 206.

Mathur, S. & Dinakarpandian, D. (2012), 'Finding disease similarity based on implicit semantic similarity', *Journal of Biomedical Informatics* **45**(2), 363–371.

Matthews, B. W. (1975), 'Comparison of the predicted and observed secondary structure of T4 phage lysozyme', *Biochimica et Biophysica Acta (BBA)-Protein Structure* **405**(2), 442–451.

Mattick, J. S. & Makunin, I. V. (2006), 'Non-coding RNA', *Human Molecular Genetics* **15**(suppl 1), R17–R29.

McGuire, S. (2016), 'World cancer report 2014. Geneva, Switzerland: World Health Organization, international agency for research on cancer, WHO Press, 2015'.

Men, K., Duan, X., He, Z., Yang, Y., Yao, S. & Wei, Y. (2017), 'CRISPR/Cas9-mediated correction of human genetic disease', *Science China Life Sciences* **60**(5), 447–457.

Moltzahn, F., Olshen, A. B., Baehner, L., Peek, A., Fong, L., Stöppler, H., Simko, J., Hilton, J. F., Carroll, P. & Blelloch, R. (2011), 'Microfluidic-based multiplex qRT-PCR identifies diagnostic and prognostic microRNA signatures in the sera of prostate cancer patients', *Cancer Research* **71**(2), 550–560.

Moore, M. J., Scheel, T. K., Luna, J. M., Park, C. Y., Fak, J. J., Nishiuchi, E., Rice, C. M. & Darnell, R. B. (2015), 'miRNA-target chimeras reveal miRNA 3'-end pairing as a major determinant of Argonaute target specificity', *Nature Communications* **6**, 8864.

Mordelet, F. & Vert, J.-P. (2014), 'A bagging SVM to learn from positive and unlabeled examples', *Pattern Recognition Letters* **37**, 201–209.

Moreno-Mateos, M. A., Vejnar, C. E., Beaudoin, J.-D., Fernandez, J. P., Mis, E. K., Khokha, M. K. & Giraldez, A. J. (2015), 'CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo', *Nature Methods* **12**(10), 982.

Najm, F. J., Strand, C., Donovan, K. F., Hegde, M., Sanson, K. R., Vaimberg, E. W., Sullender, M. E., Hartenian, E., Kalani, Z., Fusi, N. et al. (2018), 'Orthologous CRISPR–Cas9 enzymes for combinatorial genetic screens', *Nature Biotechnology* **36**(2), 179.

Ning, S., Zhang, J., Wang, P., Zhi, H., Wang, J., Liu, Y., Gao, Y., Guo, M., Yue, M., Wang, L. et al. (2016), 'Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers', *Nucleic Acids Research* **44**(D1), D980–D985.

Nishimasu, H., Ran, F. A., Hsu, P. D., Konermann, S., Shehata, S. I., Dohmae, N., Ishitani, R., Zhang, F. & Nureki, O. (2014), 'Crystal

structure of Cas9 in complex with guide RNA and target DNA', *Cell* **156**(5), 935–949.

Palazzo, A. F. & Lee, E. S. (2015), 'Non-coding RNA: what is functional and what is junk?', *Frontiers in Genetics* **6**, 2.

Pasquier, C. & Gardès, J. (2016), 'Prediction of miRNA-disease associations with a vector space model', *Scientific Reports* **6**, 27036.

Pattanayak, V., Lin, S., Guilinger, J. P., Ma, E., Doudna, J. A. & Liu, D. R. (2013), 'High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity', *Nature Biotechnology* **31**(9), 839.

Peng, R., Lin, G. & Li, J. (2016), 'Potential pitfalls of CRISPR/Cas9-mediated genome editing', *The FEBS journal* **283**(7), 1218–1231.

Petryszak, R., Keays, M., Tang, Y. A., Fonseca, N. A., Barrera, E., Burdett, T., Füllgrabe, A., Fuentes, A. M.-P., Jupp, S., Koskinen, S. et al. (2015), 'Expression Atlas updatean integrated database of gene and protein expression in humans, animals and plants', *Nucleic Acids Research* **44**(D1), D746–D752.

Platt, R. J., Chen, S., Zhou, Y., Yim, M. J., Swiech, L., Kempton, H. R., Dahlman, J. E., Parnas, O., Eisenhaure, T. M., Jovanovic, M. et al. (2014), 'CRISPR-Cas9 knockin mice for genome editing and cancer modeling', *Cell* **159**(2), 440–455.

Portin, P. & Wilkins, A. (2017), 'The evolving definition of the term "gene"', *Genetics* **205**(4), 1353–1364.

Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M. & Wain, H. (2001), 'The HUGO gene nomenclature committee (HGNC)', *Human Genetics* **109**(6), 678–680.

Pruitt, K. D., Tatusova, T. & Maglott, D. R. (2006), 'NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins', *Nucleic Acids Research* **35**(suppl_1), D61–D65.

Quek, X. C., Thomson, D. W., Maag, J. L., Bartonicek, N., Signal, B., Clark, M. B., Gloss, B. S. & Dinger, M. E. (2014), 'lncRNAdb v2. 0: expanding the reference database for functional long noncoding RNAs', *Nucleic Acids Research* **43**(D1), D168–D173.

Rahman, M. K. & Rahman, M. S. (2017), 'CRISPRpred: A flexible and efficient tool for sgRNAs on-target activity prediction in CRISPR/Cas9 systems', *PLoS ONE* **12**(8), e0181943.

Rajewsky, N. (2006), 'microRNA target predictions in animals', *Nature Genetics* **38**, S8.

Ran, F. A., Cong, L., Yan, W. X., Scott, D. A., Gootenberg, J. S., Kriz, A. J., Zetsche, B., Shalem, O., Wu, X., Makarova, K. S. et al. (2015), 'In vivo genome editing using Staphylococcus aureus Cas9', *Nature* **520**(7546), 186–191.

Ran, F. A., Hsu, P. D., Lin, C.-Y., Gootenberg, J. S., Konermann, S., Trevino, A. E., Scott, D. A., Inoue, A., Matoba, S., Zhang, Y. et al. (2013), 'Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity', *Cell* **154**(6), 1380–1389.

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A. & Zhang, F. (2013), 'Genome engineering using the CRISPR-Cas9 system', *Nature Protocols* **8**(11), 2281.

Rappaport, N., Nativ, N., Stelzer, G., Twik, M., Guan-Golan, Y., Iny Stein, T., Bahir, I., Belinky, F., Morrey, C. P., Safran, M. & Lancet, D. (2013), 'MalaCards: an integrated compendium for diseases and their

annotation', *Database* **2013**, bat018.
**URL:** *http://dx.doi.org/10.1093/database/bat018*

Ren, J., Liu, Q., Ellis, J. & Li, J. (2015), 'Positive-unlabeled learning for the prediction of conformational B-cell epitopes', *BMC Bioinformatics* **16**(18), S12.

Ren, X., Yang, Z., Xu, J., Sun, J., Mao, D., Hu, Y., Yang, S.-J., Qiao, H.-H., Wang, X., Hu, Q. et al. (2014), 'Enhanced specificity and efficiency of the CRISPR/Cas9 system with optimized sgRNA parameters in Drosophila', *Cell Reports* **9**(3), 1151–1162.

Resnik, P. et al. (1999), 'Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language', *Journal of Artificial Intelligence Research* **11**, 95–130.

Reyon, D., Tsai, S. Q., Khayter, C., Foden, J. A., Sander, J. D. & Joung, J. K. (2012), 'FLASH assembly of TALENs for high-throughput genome editing', *Nature Biotechnology* **30**(5), 460.

Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L. & Bradley, A. (2004), 'Identification of mammalian microRNA host genes and transcription units', *Genome Research* **14**(10a), 1902–1910.

Roper, J., Tammela, T., Akkad, A., Almeqdadi, M., Santos, S. B., Jacks, T. & Yilmaz, Ö. H. (2018), 'Colonoscopy-based colorectal cancer modeling in mice with CRISPR–Cas9 genome editing and organoid transplantation', *Nature Protocols* **13**(2), 217.

Saeys, Y., Inza, I. & Larrañaga, P. (2007), 'A review of feature selection techniques in bioinformatics', *Bioinformatics* **23**(19), 2507–2517.

Safavian, S. R. & Landgrebe, D. (1991), 'A survey of decision tree classifier methodology', *IEEE Transactions on Systems, Man, and Cybernetics* **21**(3), 660–674.

Sánchez, Y. & Huarte, M. (2013), 'Long non-coding RNAs: challenges for diagnosis and therapies', *Nucleic Acid Therapeutics* **23**(1), 15–20.

Schietgat, L., Vens, C., Struyf, J., Blockeel, H., Kocev, D. & Džeroski, S. (2010), 'Predicting gene function using hierarchical multi-label decision tree ensembles', *BMC Bioinformatics* **11**(1), 2.

Schliep, A., Schönhuth, A. & Steinhoff, C. (2003), 'Using hidden Markov models to analyze gene expression time course data', *Bioinformatics* **19**(suppl 1), i255–i263.

Schmitz, U., Lai, X., Winter, F., Wolkenhauer, O., Vera, J. & Gupta, S. K. (2014), 'Cooperative gene regulation by microRNA pairs and their identification using a computational workflow', *Nucleic Acids Research* **42**(12), 7539–7552.

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., Feng, G. & Kibbe, W. A. (2012), 'Disease Ontology: a backbone for disease semantic integration', *Nucleic Acids Research* **40**(D1), D940–D946.

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G. et al. (2014), 'Genome-scale CRISPR-Cas9 knockout screening in human cells', *Science* **343**(6166), 84–87.

Shen, B., Zhang, W., Zhang, J., Zhou, J., Wang, J., Chen, L., Wang, L., Hodgkins, A., Iyer, V., Huang, X. et al. (2014), 'Efficient genome modification by CRISPR-Cas9 nickase with minimal off-target effects', *Nature Methods* **11**(4), 399–402.

Shi, H., Xu, J., Zhang, G., Xu, L., Li, C., Wang, L., Zhao, Z., Jiang, W., Guo, Z. & Li, X. (2013), 'Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes', *BMC Systems Biology* **7**(1), 101.

Singh, R., Kuscu, C., Quinlan, A., Qi, Y. & Adli, M. (2015), 'Cas9-chromatin binding information enables more accurate CRISPR off-target prediction', *Nucleic Acids Research* **43**(18), e118–e118.

Stemmer, M., Thumberger, T., del Sol Keyer, M., Wittbrodt, J. & Mateo, J. L. (2015), 'CCTop: an intuitive, flexible and reliable CRISPR/Cas9 target prediction tool', *PLoS ONE* **10**(4), e0124633.

Sternberg, S. H. & Doudna, J. A. (2015), 'Expanding the biologists toolkit with CRISPR-Cas9', *Molecular Cell* **58**(4), 568–574.

Stewart, B. W. & Wild, C. (2014), 'World Cancer Report 2014. Lyon, France: International Agency for Research on Cancer', *World Health Organization* p. 630.

Sun, J., Shi, H., Wang, Z., Zhang, C., Liu, L., Wang, L., He, W., Hao, D., Liu, S. & Zhou, M. (2014), 'Inferring novel lncRNA–disease associations based on a random walk model of a lncRNA functional similarity network', *Molecular BioSystems* **10**(8), 2074–2081.

Sun, L., Luo, H., Liao, Q., Bu, D., Zhao, G., Liu, C., Liu, Y. & Zhao, Y. (2013), 'Systematic study of human long intergenic non-coding RNAs and their impact on cancer', *Science China Life Sciences* **56**(4), 324–334.

Swiech, L., Heidenreich, M., Banerjee, A., Habib, N., Li, Y., Trombetta, J., Sur, M. & Zhang, F. (2015), 'In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9', *Nature Biotechnology* **33**(1), 102–106.

Tan, M., Wu, J. & Cai, Y. (2013), 'Suppression of Wnt signaling by the miR-29 family is mediated by demethylation of WIF-1 in non-small-cell lung cancer', *Biochemical and Biophysical Research Communications* **438**(4), 673–679.

Tang, X., Lowder, L. G., Zhang, T., Malzahn, A. A., Zheng, X., Voytas, D. F., Zhong, Z., Chen, Y., Ren, Q., Li, Q. et al. (2017), 'A CRISPR–Cpf1 system for efficient genome editing and transcriptional repression in plants', *Nature Plants* **3**(3), 17018.

Thum, T., Gross, C., Fiedler, J., Fischer, T., Kissler, S., Bussen, M., Galuppo, P., Just, S., Rottbauer, W., Frantz, S. et al. (2008), 'MicroRNA-21 contributes to myocardial disease by stimulating MAP kinase signalling in fibroblasts', *Nature* **456**(7224), 980–984.

Torlay, L., Perrone-Bertolotti, M., Thomas, E. & Baciu, M. (2017), 'Machine learning–XGBoost analysis of language networks to classify patients with epilepsy', *Brain Informatics* **4**(3), 159.

Tsai, S. Q., Nguyen, N. T., Malagon-Lopez, J., Topkar, V. V., Aryee, M. J. & Joung, J. K. (2017), 'CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets', *Nature Methods* **14**(6), 607.

Tsai, S. Q., Zheng, Z., Nguyen, N. T., Liebers, M., Topkar, V. V., Thapar, V., Wyvekens, N., Khayter, C., Iafrate, A. J., Le, L. P. et al. (2015), 'GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases', *Nature Biotechnology* **33**(2), 187.

Urnov, F. D., Miller, J. C., Lee, Y.-L., Beausejour, C. M., Rock, J. M., Augustus, S., Jamieson, A. C., Porteus, M. H., Gregory, P. D. & Holmes, M. C. (2005), 'Highly efficient endogenous human gene correction using designed zinc-finger nucleases', *Nature* **435**(7042), 646.

Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S. & Gregory, P. D. (2010), 'Genome editing with engineered zinc finger nucleases', *Nature Reviews Genetics* **11**(9), 636.

Van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G. & Leunissen, J. A. (2006), 'A text-mining analysis of the human phenome', *European Journal of Human Genetics* **14**(5), 535–542.

Volders, P.-J., Helsens, K., Wang, X., Menten, B., Martens, L., Gevaert, K., Vandesompele, J. & Mestdagh, P. (2012), 'LNCipedia: a database for annotated human lncRNA transcript sequences and structures', *Nucleic Acids Research* **41**(D1), D246–D251.

Wahid, F., Shehzad, A., Khan, T. & Kim, Y. Y. (2010), 'MicroRNAs: synthesis, mechanism, function, and recent clinical trials', *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* **1803**(11), 1231–1243.

Wahlestedt, C. (2013), 'Targeting long non-coding RNA to therapeutically upregulate gene expression', *Nature Reviews Drug Discovery* **12**(6), 433–446.

Wang, H., Yang, H., Shivalila, C. S., Dawlaty, M. M., Cheng, A. W., Zhang, F. & Jaenisch, R. (2013), 'One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering', *Cell* **153**(4), 910–918.

Wang, J., Ma, R., Ma, W., Chen, J., Yang, J., Xi, Y. & Cui, Q. (2016), 'LncDisease: a sequence based bioinformatics tool for predicting lncRNA-disease associations', *Nucleic Acids Research* **44**(9), e90–e90.

Wang, K. C. & Chang, H. Y. (2011), 'Molecular mechanisms of long noncoding RNAs', *Molecular Cell* **43**(6), 904–914.

Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. (2014), 'Genetic screens in human cells using the CRISPR-Cas9 system', *Science* **343**(6166), 80–84.

Wang, X., Wang, Y., Wu, X., Wang, J., Wang, Y., Qiu, Z., Chang, T., Huang, H., Lin, R.-J. & Yee, J.-K. (2015), 'Unbiased detection of off-

target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors', *Nature Biotechnology* **33**(2), 175–178.

Wang, Y., Chen, L., Chen, B., Li, X., Kang, J., Fan, K., Hu, Y., Xu, J., Yi, L., Yang, J. et al. (2013), 'Mammalian ncRNA-disease repository: a global view of ncRNA-mediated disease network', *Cell Death & Disease* **4**(8), e765.

Ward, J. J., McGuffin, L. J., Buxton, B. F. & Jones, D. T. (2003), 'Secondary structure prediction with support vector machines', *Bioinformatics* **19**(13), 1650–1655.

Wheeler, T. J., Clements, J., Eddy, S. R., Hubley, R., Jones, T. A., Jurka, J., Smit, A. F. & Finn, R. D. (2013), 'Dfam: a database of repetitive DNA based on profile hidden Markov models', *Nucleic Acids Research* **41**(D1), D70–D82.

Wong, N., Liu, W. & Wang, X. (2015), 'WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system', *Genome Biology* **16**(1), 218.

Wu, B., Li, C., Zhang, P., Yao, Q., Wu, J., Han, J., Liao, L., Xu, Y., Lin, R., Xiao, D. et al. (2013), 'Dissection of miRNA-miRNA interaction in esophageal squamous cell carcinoma', *PLoS ONE* **8**(9), e73191.

Wu, Y., Liang, D., Wang, Y., Bai, M., Tang, W., Bao, S., Yan, Z., Li, D. & Li, J. (2013), 'Correction of a genetic disease in mouse via use of CRISPR-Cas9', *Cell Stem Cell* **13**(6), 659–662.

Wu, Y., Zhou, H., Fan, X., Zhang, Y., Zhang, M., Wang, Y., Xie, Z., Bai, M., Yin, Q., Liang, D. et al. (2015), 'Correction of a genetic disease by CRISPR-Cas9-mediated gene editing in mouse spermatogonial stem cells', *Cell Research* **25**(1), 67.

Xia, H., Ooi, L. L. P. & Hui, K. M. (2013), 'MicroRNA-216a/217-induced epithelial-mesenchymal transition targets PTEN and SMAD7

to promote drug resistance and recurrence of liver cancer', *Hepatology* **58**(2), 629–641.

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X. & Li, T. (2009), 'miRecords: an integrated resource for microRNA–target interactions', *Nucleic Acids Research* **37**(suppl 1), D105–D110.

Xiao, Y., Xu, C., Guan, J., Ping, Y., Fan, H., Li, Y., Zhao, H. & Li, X. (2012), 'Discovering dysfunction of multiple microRNAs cooperation in disease by a conserved microRNA co-expression network', *PLoS ONE* **7**(2), e32201.

Xie, B., Ding, Q., Han, H. & Wu, D. (2013), 'miRCancer: a microRNAcancer association database constructed by text mining on literature', *Bioinformatics* **29**(5), 638–644.
**URL:** *http://dx.doi.org/10.1093/bioinformatics/btt014*

Xiong, J.-S., Ding, J. & Li, Y. (2015), 'Genome-editing technologies and their potential application in horticultural crop breeding', *Horticulture Research* **2**, 15019.

Xu, H., Xiao, T., Chen, C.-H., Li, W., Meyer, C. A., Wu, Q., Wu, D., Cong, L., Zhang, F., Liu, J. S. et al. (2015), 'Sequence determinants of improved CRISPR sgRNA design', *Genome Research* **25**(8), 1147–1157.

Xu, J., Li, C.-X., Li, Y.-S., Lv, J.-Y., Ma, Y., Shao, T.-T., Xu, L.-D., Wang, Y.-Y., Du, L., Zhang, Y.-P. et al. (2011), 'MiRNA–miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features', *Nucleic Acids Research* **39**(3), 825–836.

Xu, J., Li, C.-X., Lv, J.-Y., Li, Y.-S., Xiao, Y., Shao, T.-T., Huo, X., Li, X., Zou, Y., Han, Q.-L. et al. (2011), 'Prioritizing candidate disease miRNAs by topological features in the mirna target–dysregulated network: Case study of prostate cancer', *Molecular Cancer Therapeutics* **10**(10), 1857–1866.

Xu, J., Li, Y., Li, X., Li, C., Shao, T., Bai, J., Chen, H. & Li, X. (2013), 'Dissection of the potential characteristic of miRNA–miRNA functional synergistic regulations', *Molecular BioSystems* **9**(2), 217–224.

Xuan, P., Han, K., Guo, M., Guo, Y., Li, J., Ding, J., Liu, Y., Dai, Q., Li, J., Teng, Z. et al. (2013), 'Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors', *PLoS ONE* **8**(8), e70204.

Xuan, P., Han, K., Guo, Y., Li, J., Li, X., Zhong, Y., Zhang, Z. & Ding, J. (2015), 'Prediction of potential disease-associated microRNAs based on random walk', *Bioinformatics* **31**(11), 1805–1815.

Yang, P., Humphrey, S. J., James, D. E., Yang, Y. H. & Jothi, R. (2015), 'Positive-unlabeled ensemble learning for kinase substrate prediction from dynamic phosphoproteomics data', *Bioinformatics* **32**(2), 252–259.

Yang, P., Li, X.-L., Mei, J.-P., Kwoh, C.-K. & Ng, S.-K. (2012), 'Positive-unlabeled learning for disease gene identification', *Bioinformatics* **28**(20), 2640–2647.

Yang, X., Gao, L., Guo, X., Shi, X., Wu, H., Song, F. & Wang, B. (2014), 'A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases', *PLoS ONE* **9**(1), e87797.

Yin, C., Zhang, T., Qu, X., Zhang, Y., Putatunda, R., Xiao, X., Li, F., Xiao, W., Zhao, H., Dai, S. et al. (2017), 'In Vivo Excision of HIV-1 Provirus by saCas9 and Multiplex Single-Guide RNAs in Animal Models', *Molecular Therapy* **25**(5), 1168–1186.

Yonezawa, T., Enokida, H., Yoshino, H., Hidaka, H., Yamasaki, T., Itesako, T., Seki, N. & Nakagawa, M. (2013), '461 MICRORNA-29 FAMILY AS TUMOR SUPPRESSIVE MICRORNAS IN RENAL CELL CARCINOMA: MICRORNA-29A INHIBITS CELL MIGRATION

AND INVASION TARGETING FOCAL ADHESION AND ECM PATHWAYS', *The Journal of Urology* **189**(4), e189.

Yoon, S. & De Micheli, G. (2005), 'Prediction of regulatory modules comprising microRNAs and target genes', *Bioinformatics* **21**(suppl_2), ii93–ii100.

Yu, G., Wang, L.-G., Yan, G.-R. & He, Q.-Y. (2015), 'DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis', *Bioinformatics* **31**(4), 608–609.

Yu, P.-N., Yan, M.-D., Lai, H.-C., Huang, R.-L., Chou, Y.-C., Lin, W.-C., Yeh, L.-T. & Lin, Y.-W. (2014), 'Downregulation of miR-29 contributes to cisplatin resistance of ovarian cancer cells', *International Journal of Cancer* **134**(3), 542–551.

Yu, W., Mookherjee, S., Chaitankar, V., Hiriyanna, S., Kim, J.-W., Brooks, M., Ataeijannati, Y., Sun, X., Dong, L., Li, T. et al. (2017), 'Nrl knockdown by AAV-delivered CRISPR/Cas9 prevents retinal degeneration in mice', *Nature Communications* **8**, 14716.

Zetsche, B., Gootenberg, J. S., Abudayyeh, O. O., Slaymaker, I. M., Makarova, K. S., Essletzbichler, P., Volz, S. E., Joung, J., van der Oost, J., Regev, A. et al. (2015), 'Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system', *Cell* **163**(3), 759–771.

Zhang, L., Ai, H., Chen, W., Yin, Z., Hu, H., Zhu, J., Zhao, J., Zhao, Q. & Liu, H. (2017), 'CarcinoPred-EL: Novel models for predicting the carcinogenicity of chemicals using molecular fingerprints and ensemble learning methods', *Scientific Reports* **7**(1), 2118.

Zhang, S., Zhou, J., Hu, H., Gong, H., Chen, L., Cheng, C. & Zeng, J. (2015), 'A deep learning framework for modeling structural features of RNA-binding protein targets', *Nucleic Acids Research* **44**(4), e32–e32.

Zhang, X., Zhao, X., Fiskus, W., Lin, J., Lwin, T., Rao, R., Zhang, Y., Chan, J. C., Fu, K., Marquez, V. E. et al. (2012), 'Coordinated silencing of MYC-mediated miR-29 by HDAC3 and EZH2 as a therapeutic target of histone modification in aggressive B-Cell lymphomas', *Cancer Cell* **22**(4), 506–523.

Zhao, T., Xu, J., Liu, L., Bai, J., Xu, C., Xiao, Y., Li, X. & Zhang, L. (2015), 'Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features', *Molecular BioSystems* **11**(1), 126–136.

Zhao, Y., Li, H., Fang, S., Kang, Y., Wu, W., Hao, Y., Li, Z., Bu, D., Sun, N., Zhang, M. Q. et al. (2015), 'NONCODE 2016: an informative and valuable data source of long non-coding RNAs', *Nucleic Acids Research* **44**(D1), D203–D208.

Zhou, M., Wang, X., Li, J., Hao, D., Wang, Z., Shi, H., Han, L., Zhou, H. & Sun, J. (2015), 'Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network', *Molecular BioSystems* **11**(3), 760–769.

Zhou, X., Menche, J., Barabási, A.-L. & Sharma, A. (2014), 'Human symptoms–disease network', *Nature Communications* **5**, 4212.

Zhu, S., Li, W., Liu, J., Chen, C.-H., Liao, Q., Xu, P., Xu, H., Xiao, T., Cao, Z., Peng, J. et al. (2016), 'Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR–Cas9 library', *Nature Biotechnology* **34**(12), 1279.