Graduate School of Advanced Science and Engineering
Waseda University

博士論文概要
Doctoral Thesis Synopsis

論文題目
Thesis Theme

Signal Detection and Biological Feature Extraction for High-throughput Data of N6-methyladenosine (m6A)

申請者
(Applicant Name)

Yiqian	ZHANG
張	怡倩

Department of Electrical Engineering and Bioscience,
Research on Bioinformatics

June, 2020

RNA modification is biochemical modifications of RNA and plays their roles in gene regulation at post-transcription level. So far, more than 150 types of RNA modifications have been discovered with N6-methyladenosine (m6A) as being one of the most abundant types found in nature. m6A is featured with its preferential location near 3' untranslated regions (3' UTR) and its nearby sequences mostly conforming to a certain motif, i.e., DRACH (where D = A, G or U; R = A or G; H= A, C or U) in the mammalian genome.

There are two main kinds of high-throughput sequencing technologies for mapping transcriptome-wide m6A. One is named as MeRIP-Seq(Methylated RNA immunoprecipitation sequencing, also known as m6A-seq) that was developed in 2012. The other is named as miCLIP-Seq that was developed in 2015. MeRIP-Seq detects genomic regions containing m6A sites and is economical, therefore more popular for biologists while miCLIP-Seq detects positions of m6A sites at single-base resolution.

m6A participates in essential RNA activities including alternative splicing, export, translation, and decay. During these biological processes, m6A exerts its function through interaction with several RNA binding proteins (RBPs) that can be considered as m6A-associated RBPs. There are three main types of m6A-associated RPBs, i.e. writer, eraser, and reader. m6A writer is methyltransferase including METTL3, METTL14, WTAP, RBM15/15B; m6 eraser is demethyltransferase including FTO, ALKBH5; m6A reader is proteins that can recognize m6A including YTH domain-containing proteins (YTHDF1/2/3), EIF3, FMR1. m6A writers and erasers can be considered as m6A regulators which directly regulate m6A while m6A readers can be considered as m6A effectors which participate in m6A regulatory network. These m6A-associated RBPs cooperate with each other to facilitate both temporal and spatial regulation. On the other hand, given m6A's essential roles in gene regulation, it has been found that dysfunction of m6A-associated RBPs is related to cancer progression. Therefore, the study of m6A and m6A-associated RBPs enables us to develop a better understanding of gene regulation mechanism and leads to potential therapeutic opportunities.

For m6A data analysis, I would like to be focused on two main challenges that researchers are interested in, one is the signal detection and the other is the biological feature extraction. For the signal detection, it means to apply statistical models to detect genomic regions with m6A from MeRIP-Seq data. Commonly used tools for m6A signal detection either require a long time or do not fully utilize features of RNA sequencing such as strand information which could cause ambiguous calling. In addition, with more attention on the treatment experiments (perturbation of methyltransferases) of MeRIP-Seq, biologists need intuitive evaluation on the treatment effect from comparison. Therefore, efficient and user-friendly software that can solve these tasks must be developed.

For the biological feature extraction, it means to identify biological features surrounding m6A-containing sequences using machine learning. Some tools built prediction models using random forest (RF) or support vector machine (SVM) algorithm with existing knowledge as feature input like a combination of k-mers and chemical properties, however these features are not easily interpretable, therefore it needs a prediction model for extracting meaningful biological information such as RNA binding proteins that can assist biologists in studying regulation mechanism of m6A. In that case, a deep learning model equipped with a motif (sequences recognized by certain proteins) detector is a good choice. In addition, although motifs learned from a deep learning model can help identify potential m6A-associated RBPs, the sequence-based feature has its limitation because not all the motifs of RBPs are available and sequences cannot reflect actual protein binding. Therefore, it also needs to integrate with other kinds of data like RBPs' binding data to extract biological information beyond sequence level.

The purpose of this thesis is to develop efficient and user-friendly software for detection of m6A signal and extract biological features surrounding m6A by applying statistical models and state-of-the-art machine learning methods. The main contributions of this thesis can be summarized into three aspects. 1) I developed DeepM6ASeq, a deep learning framework, to predict m6A-containing sequences and characterize biological features surrounding m6A sites. DeepM6ASeq is a sequence-based predictor that showed competitive performance of prediction, learned known m6A readers and a newly recognized one, FMR1, and also helped to visualize locations of m6A sites with a saliency map. 2) I developed MoAIMS (model-based analysis and inference of MeRIP-Seq), an efficient and easy-to-use software for analysis of MeRIP-Seq. For detection of m6A regions, MoAIMS achieves excellent speed and competitive performance compared with other tools, and provides user-friendly outputs for downstream analysis. MoAIMS also provides intuitive evaluation on treatment effects for MeRIP-Seq treatment datasets. 3) I designed an integrative computational framework for the identification of m6A-associated RBPs from reproducible m6A regions. The framework is composed of an enrichment analysis and a classification model. Utilizing the RBPs' binding data, the framework is able to identify known m6A-associated RBPs and also found some potential m6A-associated RBPs like RBM3 for mouse. Besides, it also helps infer interaction between m6A and m6A-associated RBPs including actions of reading and repelling beyond sequence level.

The thesis is composed of five chapters, which are demonstrated briefly in the followings,

[Chapter 1] Introduction presents the research background of m6A, one of the most abundant RNA modification, including its essential roles in gene regulation with m6A-associated RBPs, its detection methods of high-throughput sequencing technologies, overview of algorithms for m6A data analysis. Next, research objectives are proposed and the contribution of the thesis is described. Finally, the organization of the thesis is demonstrated.

[Chapter 2] DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning presents the work on m6A biological features extraction at sequence level. Existing tools can predict m6A at single-base resolution, however, the features they used for prediction such as k-mers and chemical properties are less interpretable. It needs a model to extract meaningful biological information like RNA binding proteins in the m6A-containing sequences in order to provide more insights for the biologists, therefore I implemented a deep learning framework, named DeepM6ASeq, to predict m6A-containing sequences and characterize surrounding biological features based on miCLIP-Seq data, which detects m6A sites at single-base resolution. DeepM6ASeq showed better performance as compared to other machine learning classifiers. Moreover, an independent test on MeRIP-Seq data, which identifies m6A-containing genomic regions, revealed that our model is competitive in predicting m6A-containing sequences. DeepM6ASeq utilized the convolutional neural network (CNN) layer as a motif detector and identified known m6A readers. Notably, DeepM6ASeq also identifies a newly recognized m6A reader: FMR1. Besides, I found that a saliency map in the deep learning model could be utilized to visualize locations of m6A sites.

[Chapter 3] MoAIMS: efficient software for detection of enriched regions of MeRIP-Seq presents the work on m6A signal detection of MeRIP-Seq(Methylated RNA immunoprecipitation sequencing). MeRIP-Seq is an economical and popular high-throughput sequencing method for studying m6A. The signal detection is a main challenge for data analysis of MeRIP-Seq, however current tools either require a long time or do not fully utilize features of RNA sequencing such as strand information which could cause ambiguous calling. On the other hand, with more attention on the treatment experiments of MeRIP-Seq, biologists need intuitive evaluation on the treatment effect from comparison. Therefore, efficient and user-friendly software that can solve these tasks must be developed. I developed a software named "model-based analysis and inference of MeRIP-Seq (MoAIMS)" to detect enriched regions of MeRIP-Seq and infer signal proportion based on a mixture negative-binomial model. MoAIMS is designed for transcriptome immunoprecipitation sequencing experiments; therefore, it is compatible with different RNA sequencing protocols. MoAIMS offers excellent processing speed (nearly ten times faster) and competitive performance on motif occurrence in the m6A regions and the overlapping percentage with miCLIP-Seq data when compared with other tools. Furthermore, signal proportion inferred from MoAIMS showed a decreasing trend for m6A treatment dataset (perturbation of m6A methyltransferases) compared with m6A wild-type dataset, which is consistent with experimental observations, suggesting that the signal proportion can be used as an intuitive indicator of treatment effect.

[Chapter 4] Identification of m6A-associated RBPs using an integrative computational framework presents the work on m6A biological features extraction at protein-binding level. Existing tools for extracting m6A biological features are sequence-based ones which are limited because not all the motifs of RBPs are available and sequences cannot reflect actual protein binding, therefore in this study I designed an integrative computational framework to extract m6A biological features, i.e. m6A-associated RBPs, utilizing RBP's binding data. I identified reproducible m6A regions from independent studies in certain cell lines and then utilized RBPs' binding data of the same cell line to identify m6A-associated RBPs. The computational framework is composed of an enrichment analysis and a classification model. The enrichment analysis identified known m6A-associated RBPs including YTH domain-containing proteins; it also identified a potential m6A-associated RBP, RBM3, for mouse. I observed a significant correlation for the identified m6A-associated RBPs at the protein expression level rather than the gene expression. In addition, I built a Random Forest classification model for the reproducible m6A regions using the information of RBPs' binding. The RBP-based predictor not only demonstrated competitive performance compared with sequence-based ones and but also helped identify m6A-repelled RBP. These results suggested that the framework enabled us to infer interaction between m6A and m6A-associated RBPs beyond sequence level when utilizing RBPs' binding data.

[Chapter 5] General conclusions and future work presents general conclusion for the thesis and future work. The future work has main three parts, including model improvement for the signal detection, algorithms and data worth investigating for biological feature extraction and application in other RNA modifications.

In conclusion, the key contribution of the thesis is to extract meaningful biological features like m6A-associated RBPs by applying mathematical models in the analysis of m6A high-throughput data so that it could help biologists develop more insights into regulation mechanism of m6A.

No.1

早稲田大学　博士（工学）　学位申請　研究業績書

(List of research achievements for application of doctorate (Dr. of Engineering), Waseda University)

氏 名(ZHANG, Yiqian)　　　　　　　　印(seal or signature　　　　　　)

（As of May, 2020）

種 類 別 (By Type)	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）(theme, journal name, date & year of publication, name of authors inc. yourself)
Journal	O [1] **Yiqian Zhang** and Michiaki Hamada, DeepM6ASeq: Prediction and Characterization of m6A-containing Sequences using Deep Learning, BMC Bioinformatics. 2018 Dec 31;19(Suppl 19):524. O [2] **Yiqian Zhang** and Michiaki Hamada, MoAIMS: Efficient Software for Detection of Enriched Regions of MeRIP-Seq, BMC Bioinformatics. 2020 Mar 14;21(1):103.

早稲田大学　博士（工学）　学位申請　研究業績書

(List of research achievements for application of doctorate (Dr. of Engineering), Waseda University)

種 類 別 By Type	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む） (theme, journal name, date & year of publication, name of authors inc. yourself)
International Conferences	O [1] **Yiqian Zhang** and Michiaki Hamada, DeepM6ASeq: Prediction and Characterization of m6A-containing Sequences using Deep Learning, Genome Informatics Workshop (GIW), Dec. 03-05, 2018, Kunming, China (Talk).

早稲田大学　博士（工学）　学位申請　研究業績書

(List of research achievements for application of doctorate (Dr. of Engineering), Waseda University)

種 類 別 By Type	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む） (theme, journal name, date & year of publication, name of authors inc. yourself)
Domestic Conferences	O [1] **Yiqian Zhang** and Michiaki Hamada, Prediction of mRNA m6A sites in the mammalian genome, 第 6 回生命医薬情報学連合大会(IIBMP), Sep. 27-29, 2017, 札幌 (Poster). O [2] **Yiqian Zhang** and Michiaki Hamada, Prediction and Characterization of m6A-contained Sequences using Deep Neural Network, バイオ情報学研究会(IPSJ-BIO), Jun. 13-15, 2018, 沖縄 (Talk). O [3] **Yiqian Zhang** and Michiaki Hamada, DeepM6ASeq: Prediction and Characterization of m6A-contained Sequences using Deep Neural Networks, 第 20 回日本 RNA 学会年会, Jul. 09-11, 2018, 大阪 (Poster). O [4] **Yiqian Zhang** and Michiaki Hamada, Developing an efficient software for MeRIP-Seq signal detection，第 3 回 Tokyo Bioinformatics Meeting 研究会，Aug. 29, 2019, 東京 (Talk). O [5] **Yiqian Zhang** and Michiaki Hamada, Developing efficient software for detection of enriched regions of MeRIP-Seq, 第 42 回日本分子生物学会年会, Dec. 03-06, 2019, 福岡 (Poster).