Transcriptional and Post-transcriptional Regulation of Gene Expression

2016

Jun Ding
University of Central Florida

TRANSCRIPTIONAL AND POST-TRANSCRIPTIONAL REGULATION OF GENE EXPRESSION

by

JUN DING

B.S Hefei University of Technology, 2007
MS. University of Science and Technology of China, 2010
MS. University of Central Florida, 2012

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Department of Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Spring Term
2016

Major Professor: Haiyan Nancy Hu

ABSTRACT

Regulation of gene expression includes a variety of mechanisms to increase or decrease specific gene products. Gene expression can be regulated at any stage from transcription to post-transcription and it's essential to almost all living organisms, as it increases the versatility and adaptability by allowing the cell to express the needed proteins.

In this dissertation, we comprehensively studied the gene regulation from both transcriptional and post-transcriptional points of view. Transcriptional regulation is by which cells regulate the transcription from DNA to RNA, thereby directing gene activity. Transcriptional factors (TFs) play a very important role in transcriptional regulation and they are proteins that bind to specific DNA sequences (regulatory elements) to regulate the gene expression. Current studies on TF binding are still very limited and thus, it leaves much to be improved on understanding the TF binding mechanism. To fill this gap, we proposed a variety of computational methods for predicting TF binding elements, which have been proved to be more efficient and accurate compared with other existing tools such as DREME and RSAT peaks-motif. On the other hand, studying only the transcriptional gene regulation is not enough for a comprehensive understanding. Therefore, we also studied the gene regulation at the post-transcriptional level. MicroRNAs (miRNAs) are believed to post-transcriptionally regulate the expression of thousands of target mRNAs, yet the miRNA binding mechanism is still not well understood. In this dissertation, we explored both the traditional and novel features of miRNA-binding and

proposed several computational models for miRNA target prediction. The developed tools outperformed the traditional microRNA target prediction methods (.e.g miRanda and TargetScan) in terms of prediction accuracy (precision, recall) and time efficiency.

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my advisor Dr. Haiyan Hu, who continually and convincingly tell me that a good attitude is very important to research and work. That's extremely helpful for my Ph.D. studies. Without her guidance and persistent help, this dissertation would not have been possible.

I would also like to thank my co-advisor Dr. Xiaoman Li, who has the attitude and the substance of a genius. I have been learning so much from him in the past 5 years and I am really grateful for the great help he offered to my research.

I also want to take this opportunity to thank my committee members Dr. Shaojie Zhang and Dr. Yier Jin for their dedication. Their questions and suggestions to my dissertation are very helpful and illuminating.

Lastly, I would also like to thank my wife, parents, and my daughter. They were always supporting me and encouraging me with their best wishes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Regulation of gene expression includes a variety of mechanisms to increase or decrease specific gene products. These products are often proteins, but for those non-protein coding genes, the products are functional RNAs, such as transfer RNA (tRNA) and small nuclear RNA (snRNA). Gene expression can be regulated at any stage from transcription to post-transcription. The regulation of gene expression is essential to almost all living organisms, including viruses, prokaryotes, and eukaryotes, as it increases the versatility and adaptability by allowing the cell to express the needed proteins. Abnormal gene expression usually leads to various diseases. Diagnosing and curing these diseases demands better understanding of the gene regulation mechanisms. Besides, better understanding of gene regulation mechanisms is also very helpful to bio-engineering field, as we might be able to significantly improve the bio-production with better regulation knowledge of specific genes. The gene regulation is of vital importance, yet our understanding is still very limited. In this dissertation, we studied the regulation of the gene expression from both transcriptional and post-transcriptional points of view.

1.1 Studies of Gene Regulation at Transcriptional Level

For the transcriptional regulation of the gene expression, direct interaction with DNA is the simplest and most direct way to regulate the gene expression. It's estimated that there are 1500 transcription factors (TFs) that plays a key role in gene regulation by directly binding to the DNA alone or through a protein complex (Vaquerizas, Kummerfeld et al. 2009). The mechanism of TF binding is not completely understood. One challenge in improving the understanding is the discovery of sequence motifs, which TFs used to recognize its binding targets. A DNA motif is defined as a DNA sequence

pattern that has some biological significance as being DNA binding sites for a regulatory protein (e.g. transcription factor). Normally, the pattern is short (5-20 bps) and is known to recur in different genes or multiple times within a gene (Lescot, Déhais et al. 2002). A variety of DNA motif finding algorithms have been developed: MEME (Bailey, Williams et al. 2006), Bioprospector (Liu, Brutlag et al. 2001), MDScan (Liu, Brutlag et al. 2002), Phylocon (Wang and Stormo 2003), etc. Those traditional methods are mainly based on the TF binding features like: over-representation, conservation, etc. In this dissertation, we proposed a traditional type of method called MERCED, which take the mutation/substitution into the consideration of conservation. The comparison with other conservation based algorithms, such as Phylocon, shows the superior performance of MERCED.

With the recent advance of Next Generation sequencing technique (NGS), Chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing (ChIP-Seq) is playing a more and more important role in finding TF binding sites. The actual binding location is typically within ~50bp of the predicted location indicated by ChIP-Seq data (Wilbanks and Facciotti 2010). A typical TF ChIP-Seq experiment yields up to thousands of predicted binding locations (TF peaks). Most existing traditional motif finding algorithms such as MEME cannot handle such large datasets. Recently, there are a few motif finding algorithms, which can be applied to the ChIP-seq data, have been developed, such as DREME (Bailey 2011) and RSAT: Peaks-motif (Thomas-Chollier, Herrmann et al. 2012). Several computational methods identify motifs in top ChIP-seq peak regions (Jothi, Cuddapah et al. 2008, Valouev, Johnson et al. 2008). Such a type of approaches is likely to miss many potential motifs since TFBSs of cofactors may only occur in some ChIP-seq peaks (Bailey 2011). A few methods attempt to identify TFBSs and motifs in all peak regions, by using

known motifs to scan (extended) ChIP-seq peak regions to identify significantly co-occurring motifs (Sun, Guns et al. 2012, Ding, Cai et al. 2013). This type of approaches has achieved certain success in identifying motifs of underrepresented cofactors (Ding, Cai et al. 2013). Since current knowledge of known motifs is still limited, many TFBSs and motifs are likely missed by this type of methods. There are also methods for de novo discovery of TFBSs and motifs in all peak regions from a ChIP-seq experiment (Hu, Yu et al. 2010, Kulakovskiy, Boeva et al. 2010, Bailey 2011, Thomas-Chollier, Herrmann et al. 2012). Note that TFBSs of certain cofactors may only occur in a small number of peaks (Bailey 2011, Stamatoyannopoulos 2012). Motifs of these cofactors may thus be statistically insignificant individually. The existing de novo motif discovery methods may thus miss motifs and TFBSs of many cofactors.

Here we proposed a novel computational approach SIOMICS (systematic identification of motifs in Chip-seq data) (Ding, Hu et al. 2014) for de novo discovery of motifs and TFBSs from all peak regions of a ChIP-seq experiment. Instead of considering individual motifs separately for motif discovery, SIOMICS simultaneously considers motif modules, i.e., combinations of any number of motifs that co-occur in at least a predefined number of peak regions and have p-value of statistical significance smaller than a given threshold. Instead of considering only motifs that are significantly overrepresented in ChIP-seq peak regions, SIOMICS takes both overrepresented and non-overrepresented motifs into account. Tested on 13 ChIP-seq datasets, SIOMICS identified many known motifs, new motifs, and their TFBSs. Tested on 13 simulated random datasets that were obtained by permuting the experimental sequence data, SIOMICS did not predict any false motif. Compared with two recent methods, DREME (Bailey 2011) and Peak-motifs (Thomas-Chollier,

Herrmann et al. 2012), SIOMICS identified more known cofactor motifs in ChIP-seq datasets and the same or fewer false motifs in random datasets and had a comparable or better time efficiency.

Although SIOMICS has good running efficiency and performance, it can only predict length-specific motifs, which is one of its major disadvantages. In order to overcome this disadvantage, we proposed an improved version of SIOMICS, we call it SIOMICS-Extension(Ding, Dhillon et al. 2015). The major improvement of the SIOMICS tool in the extended version is its ability to output motifs of different lengths. In practice, the common motif length varies from 5 to 15 base pairs (Wingender, Dietze et al. 1996). The initial SIOMICS tools may thus cause certain problems by requiring the predicted motif length be always *w* base pairs long. To address this issue, SIOMICS_Extension first considers extending these motifs from the left side, or the right side, or both sides.

After extending the motifs, there may be some motifs left unchanged. For these motifs, SIOMICS_Extension considers reducing their motif lengths. SIOMICS_Extension compares these unchanged motifs to see whether two unchanged motifs are similar. Moreover, SIOMICS_Extension checks whether these similar motifs form motif modules with a common group of other motifs. If both criteria are satisfied, SIOMICS_Extension considers whether to merge these similar motifs and represent them with the similar portion.

1.2 Studies of Gene Regulation at Post-Transcriptional Level

Post-transcriptional regulation means the control of gene expression at the RNA level, which is between the transcription and the translation of a gene. The majority of gene regulation studies to date was mainly focused on transcriptional regulation mechanism, but the importance of post-transcriptional gene expression in eukaryotes is becoming increasingly clear. MicroRNAs (miRNAs)

play a critical role in gene regulation (Bartel 2004, Bartel 2009) and it's reported that microRNAs appear to post-transcriptionally regulate the expression of more than 60% of protein coding genes of human genome (Friedman, Farh et al. 2009). In this dissertation, we focused our study of post-transcriptional regulation on microRNAs. MiRNA is a family of small non-coding RNAs of ~22 nucleotides long. They can bind mRNAs at 5' untranslated regions (UTRs), coding sequences (CDSs), and 3' UTRs. The binding is traditionally thought to be through the base-pairing of the seed regions in miRNAs with the partially complementary sequences in the target mRNAs (Bartel 2009). The seed region refers to the 5' end miRNAs from the position 2 to the position 7 (Lewis, Shih et al. 2003, Lewis, Burge et al. 2005). Depending on the pairing quality, the miRNA target sites are classified into two categories: canonical sites and non-canonical sites. The former is the target sites that are perfect complementary to the seed regions, while the latter is the target sites that have imperfect seed complementarity (G:U wobbles or mismatches). With the advance of biotechnology, currently, it is commonly accepted that base-pairing can involve both seed regions and outside of the seed regions (Hafner, Landthaler et al. 2010, Helwak, Kudla et al. 2013, Wang 2014). That is, other types of target sites exist in addition to the canonical and non-canonical target sites. We define target sites other than the canonical sites as unconventional sites. Regardless of the types of target sites, the binding of miRNAs to their target mRNAs during diverse cellular processes can degrade the target mRNAs, and/or repress the translation of the target mRNAs to proteins (Bartel 2004, Bartel 2009, Wang, Li et al. 2011). Due to such a pivotal role in gene regulation, it is critical to study miRNAs and their target sites.

MiRNAs often form modules to regulate their target mRNAs (Doench and Sharp 2004, Vella, Choi et al. 2004, Krek, Grun et al. 2005, Saetrom, Heale et al. 2007, Wu, Huang et al. 2010). Informally, a

miRNA module is a group of miRNAs that bind a common set of mRNAs under the same experimental condition. Several studies related to miRNA modules have been published in the past decade (Krek, Grun et al. 2005, Saetrom, Heale et al. 2007, Wu, Huang et al. 2010, Jayaswal, Lutherborrow et al. 2011, Zhang, Li et al. 2011, Bryan, Terrile et al. 2013). Among them, Krek et al. considered the co-occurrence of multiple miRNA target sites in 3' UTRs to score mRNAs as potential miRNA targets and predicted miRNA targets (Krek, Grun et al. 2005). Saetrom et al. investigated the preferred distances of target sites of the same miRNAs in 3' UTRs (Saetrom, Heale et al. 2007). Other studies predicted miRNA modules by harnessing the predicted miRNA target sites in 3' UTRs and the co-expression relationship of target mRNAs of the same miRNAs (Jayaswal, Lutherborrow et al. 2011, Zhang, Li et al. 2011, Bryan, Terrile et al. 2013).

Despite the significant predictions and discoveries from these studies, our understanding of miRNA modules is still rudimentary. To our knowledge, all published studies on miRNA modules so far are based on the computationally predicted miRNA target sites in 3' UTRs. However, 3' UTRs only account for a small portion of the potential miRNA target site residing regions (Hafner, Landthaler et al. 2010, Helwak, Kudla et al. 2013). Moreover, even the most well-known target site prediction methods currently produce a significant fraction of false positive predictions (Witkos, Koscianska et al. 2011). In addition, in defining miRNA modules, rarely does a study require the higher down-regulation of target gene expression by miRNA modules than that by subsets of miRNAs contained in the modules (Wu, Huang et al. 2010). Therefore, although we have gained basic insight into miRNA modules from previous studies, our understanding of miRNA modules may be biased and limited. More importantly, several key questions have not been addressed in previous studies. For instance, is there any difference between the target sites bound by miRNA modules and those bound

by individual miRNAs? What is the preferred distance range of the adjacent target sites of different miRNAs in a miRNA module? And so on. Many aspects of miRNA modules and their target sites remain elusive.

To address these questions, we employed experimentally determined instead of computationally predicted target sites to study miRNA modules. We used the recently generated high-throughput data from the crosslinking, ligation, and sequencing of hybrids (CLASH) (Helwak, Kudla et al. 2013). The CLASH data provides an unprecedented opportunity to study miRNA modules because it provides information about which miRNAs bind which short mRNA regions under a given experimental condition.

Besides, we also developed a new approach for miRNA target site prediction called TarPmiR (Target Prediction for miRNAs). TarPmiR applies a random-forest-based approach to integrate six conventional features and seven new features to predict miRNA target sites. These features were learned from the only CLASH dataset in mammal that is made publically available by Helwak, et al. (Helwak, Kudla et al. 2013). By cross-validation, we showed that TarPmiR had an average recall of 0.543 and an average precision of 0.181. Tested on three independent datasets, including two human PAR-CLIP datasets and one mouse HITS-CLIP dataset, we demonstrated that TarPmiR identified more than 74.2 % of known miRNA target sites in each dataset. Compared with three existing approaches, we found that TarPmiR is superior to existing approaches, in terms of both higher recall and higher precision. The TarPmiR method is implemented in a python package, which is freely available at http://hulab.ucf.edu/research/projects/miRNA/TarPmiR/.

1.3 Overview of the Dissertation

In summary, we studied the gene regulation at both transcriptional and post-transcriptional levels. Transcription Factors (TFs) play a very important role in transcriptional gene regulation by directly binding the DNA sequence, yet the TF binding mechanism is not completely clear. In this work, we proposed a sets of computational methods, which can systematically predict the TF and cofactor binding on a whole genome scale. What's more importantly, those methods can be applied to the large scale TF ChIP-Seq data, which provides much better resolution for finding TF binding sites. MicroRNAs (miRNAs) play a critical role in Post-transcriptional gene regulation. It's reported that microRNAs appear to post-transcriptionally regulate the expression of more than 60% of protein-coding genes of human genome (Friedman, Farh et al. 2009). Thereby we focused our study of post-transcriptional regulation on microRNAs and we proposed several computational methods on microRNA binding prediction. The remaining of this dissertation will be organized as follows:

(1) In Chapter 2, I will present our studies on transcriptional regulation: we proposed 3 methods on predicting TF binding: MERCED, SIOMICS, and SIOMICS2.

(2) In Chapter 3, I will present our studies on post-transcriptional regulation. We proposed 2 computational frameworks on miRNA binding target: miRModule and TarPmiR. We also analyzed the miRNA binding features, especially those miRNAs in modules.

(3) In Chapter 4, I will conclude the gene regulation studies in this dissertation and a few on-going projects will be introduced. It's expected that they can further improve the performance of TF binding and miRNA target prediction, which will also advance our understanding of gene regulation mechanisms.

CHAPTER 2: TRANSCRIPTIONAL REGULATION –TF BINDING

2.1 MERCED: Systematic Discovery of Cis-Regulatory Elements in *Chlamydomonas reinhardtii* Genome Using Comparative Genomics

2.1.1 Background

Chlamydomonas reinhardtii (*C. reinhardtii*) is a member of single celled green algae that diverged from the streptophytes approximately one billion years ago. Being comprised of multiple mitochondria, two anterior flagella and a chloroplast, *C. reinhardtii* serves as an outstanding microalgae model organism, especially for analyzing eukaryotic chloroplast biology, action of flagella and basal bodies and many biological pathways such as circadian rhythms, cell cycle control and plant respiration etc. (Rochaix 2004, Wemmer and Marshall 2004, Bisova, Krylov et al. 2005, Cardol, Gonzalez-Halphen et al. 2005, Mittag, Kiaulehn et al. 2005). As a photosynthetic microalgae species, *C. reinhardtii* has also shown its potential in biofuel generation (Grossman, Harris et al. 2003, Beckmann, Lehr et al. 2009, Langner, Jakob et al. 2009, Nguyen, Choi et al. 2009). However, since cellular processes in general are coordinated by transcriptional regulation of functionally related genes, further exploitation of *C. reinhardtii* as a model system to elucidate various molecular mechanisms requires systematic study of gene regulation (Li, Han et al. , Sun, Liu et al. , Bohne and Linden 2002). Genome-scale study of gene regulation in *C. reinhardtii* is currently in the very early stages. For example, Cis-regulatory elements (CREs) are genomic DNA segments that play important roles in gene regulation by modulating gene activities through their interaction with RNAs or regulatory proteins called *Transcription Factors (TFs)*. The discovery and functional annotation of CREs is thus one of the immediate next steps toward global understanding of gene regulatory mechanisms in *C. reinhardtii* (Wang, Haberer et al. 2009). However, there are less than a dozen CREs annotated in *C. reinhardtii* as of yet, even though many TFs have been collected and deposited

in a number of databases (Wingender, Dietze et al. 1996, Higo, Ugawa et al. 1999, Rombauts, Dehais et al. 1999, Portales-Casamar, Thongjuea et al. 2010). The correspondences between CREs and regulatory proteins remain largely unknown.

Unprecedented large-scale microalgae genomic data has now become available, which makes it possible to perform the genome-wide identification of CREs in *C. reinhardtii*. For example, the complete nuclear genome sequence of *C. reinhardtii* was published in 2007, and around 15, 000 protein-coding genes were predicted (Merchant, Prochnik et al. 2007). Additionally, another green algal species *Volvox carteri* that is close to *C. reinhardtii* in evolution has been recently sequenced, which provides a great opportunity to study the *C. reinhardtii* genome by comparative genomics (Prochnik, Umen et al. 2010). In addition to the genomic sequence data, there is a large supply of expression data available as cDNA libraries, expressed sequence tags (ESTs), microarray expression measurements, and RNA sequencing reads (Eberhard, Jain et al. 2006, Miller, Wu et al. 2010, Castruita, Casero et al. 2011, Fischer, Ledford et al. 2012, Urzica, Adler et al. 2012). Integrative analysis of DNA sequence data and mRNA expression measurements have been shown successful in gene regulatory network studies at the whole-genome level (Segal, Shapira et al. 2003, Wang and Stormo 2003).

In this paper, we present the first genome-wide computational discovery of CREs in *C. reinhardtii* using a comparative-genomics-based approach named MERCED, short for "Modeling Evolution Rate across species for Cis-regulatory Element Discovery". MERCED searches for CREs through uncovering *CRE motifs*, i.e., common patterns of CREs that can be bound by the same TFs. By simultaneously considering multiple properties of motifs including overrepresentation, co-occurrence

and evolutionary conservation, MERCED is able to reduce false positive predictions significantly. Most importantly, considering species-divergence-time when evaluating evolutionary conservation leads to better incorporation of the evolutionary-information of individual DNA segments. By comparatively integrating *V. carteri* and *C. reinhardtii* genome information, MERCED discovered 66530 CREs, corresponding to 317 CRE motifs. 164 (51.7%) of these CRE motifs tend to frequently co-occur in regulatory sequences flanking the transcription start sites. The existence of many such frequently co-occurring motifs, named as *motif combinations*, indicates the potential of these motifs to coordinately regulate target genes. Many of these identified motifs and motif combinations are consistent with experimentally verified motifs from public databases. Further integration of gene transcriptional profiles and gene-annotation data resources also provide multiple functional evidences supporting the discovery. The motif predictions generated from this study and the accompanying software tool MERCED have been deposited into our web-accessible database, which will be useful to experimental biologists interested in genes regulation in algae species.

2.1.2 Materials and Methods

2.1.2.1 Protein, DNA Sequences and Orthologous Gene Pairs in C. reinhardtii and V. carteri

We defined *promoter sequence* of a gene as the upstream 1000 base pair (1kbp) sequence relative to the translational start site of this gene. All genes and their promoter sequences in *C. reinhardtii* and *V. carteri* were downloaded from the Phytozome database (http://www.phytozome.net/). The JGI v4.3 and the JGI v1.0 release were used for the two species, respectively. In total, we obtained 17116 genes in *C. reinhardtii* and 14546 genes in *V. carteri*. Repeats in the upstream sequences were then masked using RepeatMasker (http://www.repeatmasker.org/). We also obtained all proteins sequences in the two species from the Phytozome database. The PSI-BLAST program (Altschul,

Madden et al. 1997) was then applied to these sequences to find the reciprocal best hits between proteins in the two species. In total, we defined 8742 orthologous pairs based on reciprocal best hits (E-value ≤ 1E-10) in *C. reinhardtii* and *V. carteri*.

2.1.2.2 Gene Expression Data and Co-expressed Clusters

We selected four gene expression datasets in *C. reinhardtii* from the Gene Expression Omnibus (GEO) database requiring each of them contains at least fifteen samples (Edgar, Domrachev et al. 2002). The datasets are GSE20860, GSE20861, GSE30646 and GSE30648. We then calculated the pair-wise Pearson's correlation for every pair of genes in each dataset. These gene correlations were then used to obtain co-expressed gene sets by applying a hierarchical clustering algorithm with the average linkage (Sokal and Michener 1985). In total, we obtained 437 gene sets, each of which contains more than three genes and has gene correlation no less than 0.6.

2.1.2.3 Model the Neutral Evolution Rate of Nucleotide Substitution

We constructed a substitution matrix to describe the neutral evolution of nucleotides between *C. reinhardtii* and *V. carteri*. The construction is based on the fourfold degenerate sites in orthologous proteins in the two algae species. We first aligned each pair of orthologous proteins by using the widely-used protein-alignment software MUSCLE (Edgar 2004). We then obtained all aligned fourfold degenerate sites with the same amino acid in the alignments. A fourfold degenerate site is a position in a codon where all nucleotide substitutions at this site are synonymous. For each of the positions containing fourfold degenerate sites, we obtained their corresponding nucleotides and counted how many times a given type of nucleotides in *C. reinhardtii* corresponds to another given type of nucleotides in *V. carteri*. In the end, we obtained a four by four substitution matrix *S* involving the four types of nucleotides A, C, G and T, as illustrated in Figure 2A.

2.1.2.4 Discovery of Conserved k-mers in C. reinhardtii

To identify k-mers (k-base-pair-long DNA segment) in *C. reinhardtii* that are conserved between *C. reinhardtii* and *V. carteri*, we first defined *conserved-to-be (CTB) k-mer pairs*. A CTB k-mer pair contains two k-mers, one from *C. reinhardtii* and the other from *V. carteri*, which are likely to be conserved between these two species. To find out all CTB k-mer pairs in *C. reinhardtii*, we obtained the expected number of mismatches *m* between any k-mer μ in *C. reinhardtii* and any k-mer in μ's corresponding orthologous sequence in *V. carteri*. Here, *m* is calculated as $m = k\sum_{i=A}^{T} p_i(1 - S_{ii})$, where, p_i is the probability that the i-th type of nucleotides in *C. reinhardtii* and S_{ii} is the probability that the i-th type of nucleotides in *C. reinhardtii* correspond to the i-th type of nucleotides in *V. carteri*. We defined two k-mers as a CTB k-mer pair if the observed mismatch between them is smaller than *m*. With all the CTB k-mer pairs in the two species defined, we then selected k-mers from these CTB k-mer pairs as conserved k-mers by their statistical significance (p-value) in terms of evolutionary conservation. To calculate the conservation p-value of a CTB k-mer pair that contains a specific k-mer, one needs to compare all the 4^k different k-mers in *V. carteri* with this k-mer in *C. reinhardtii*. However, with the large number of k-mers in *C. reinhardtii*, direct enumerating all the 4^k different k-mers in *V. carteri* is time-consuming. Therefore, we exploited the mathematics concept of generating function to calculate the p-value. The generating function is a representation of the probability mass function of a random variable. For a discrete random variable *X*, its generating function is defined as $f(t) = \sum_{i=1}^{n} p_i t^{a_i}$, where $\Pr(X = a_i) = p_i$ and *n* is the number of different values X can take. For every k-mer μ in *C. reinhardtii* that has similar k-mers in μ's corresponding orthologous sequence in *V. carteri*, say $a_1 a_2 ... a_k$, we can define a generating function of the k-mer in *C. reinhardtii*, as $f(t, a_1 a_2 ... a_k) = \sum_{i=1}^{4^k} c_i t^{score_i}$, where c_i is the probability of the i-th k-mer occurs

in the upstream 1000 base pair sequences in *V. carteri*, and $score_i$ is the substitution score of the k-mer $a_1a_2...a_k$ in *C. reinhardtii* by the i-th k-mer in *V. carteri*. The generating function can be efficiently computed as in previous studies (Staden 1989, Huang, Kao et al. 2004). The statistical significance for the evolutionary conservation between $a_1a_2...a_k$ and one of its similar k-mer in *V. carteri* can then be calculated as p-value = $\sum_{score_i \geq alpha} c_i$, where *alpha* is the substitution score of this k-mer pair. We claimed a k-mer pair as conserved if its Bonferroni corrected p-value ≤ 0.05.

2.1.2.5 Prediction of Motifs and Motif Combinations

To predict CRE motifs from the conserved k-mers identified in *C. reinhardtii*, we first applied the hierarchical clustering algorithm (average linkage) (Sokal and Michener 1985) to cluster the conserved k-mers. Since conserved k-mers forming a cluster are likely to have the same underlying pattern and thus correspond to the same motif, we predicted a motif for each cluster. Each predicted motif is represented by a PWM (Stormo and Hartzell 1989). In the end, we obtained 66530 conserved CREs and 317 underlying 8-mer motifs. With the predicted motifs, we further identified motif combinations by finding motifs frequently co-occurring in a large number of regulatory sequences in *C. reinhardtii* using our previously developed method (Cai, Hou et al. 2010). The statistical significance of each group of frequently co-occurring motifs was determined by Poisson clumping heuristic (Cai, Hou et al. 2010). The groups of motifs with sufficient significance (FDR < 0.05) were predicted as motif combinations.

2.1.2.6 Function Study of the Predicted Motif Combinations

To see whether motif combinations are associated with specific functions, we investigated the overlap between the target genes of co-occurring motifs and functional gene sets. The functional gene sets are defined as sets of genes with specific Gene Ontology (GO) function annotation

(Ashburner, Ball et al. 2000), or set of genes with correlated transcription expression since correlated expression often implicates similar function (Altman and Raychaudhuri 2001, Ideker, Thorsson et al. 2001). The statistical significance of any overlap between target genes of a motif combination and a given functional gene set is calculated as follows. Let S be the set of all the N genes in a genome, S_1 be a predicted target gene set of a motif combination and S_2 be a given functional gene set, and assume the number of genes in the intersection of the three sets S, S_1, and S_2 is n, M and m respectively. Then the p-value of the overlap of the set S_1 and the set S_2 can be estimated based on the hyper-geometric test: $pvalue = \sum_{i=m}^{\min(n,M)} \frac{C(M,i)C(N-M,n-i)}{C(N,n)}$, where $C(x,y)$ is the combinatorial number of choosing y items out of x items. From such obtained p-values, we then calculate q-values based on the Q-Value software to estimate the statistical significance of the overlap (Storey and Tibshirani 2003). All the motif combinations with sufficient statistical significance (FDR < 0.05) in terms of overlapping with known functional gene sets are reported as functional motif combinations.

2.1.3 Results

2.1.3.1 Genome-scale discovery of CRE motifs in C. reinhardtii

Existing CRE-motif-finding methods often assumes that a bona fide CRE motif must be overrepresented in the input sequences, i.e., the overrepresentation property of motifs (Stormo and Hartzell 1989, Lawrence, Altschul et al. 1993, Bailey and Elkan 1994). However, because of the degenerative nature of motifs, overrepresentation alone is often not enough to distinguish true motifs from random patterns formed by DNA segments (Blanchette and Tompa 2002, Wang and Stormo 2003). To improve sensitivity and specificity of the CRE motif discovery, dozens of motif-finding methods have been developed to exploit the co-occurrence property of motifs, i.e., multiple CREs are often co-occurring in short regions (Frith, Hansen et al. 2001, Zhou and Wong 2004, Gupta and Liu 2005, Hu, Hu et al. 2008). Alternative methods require evolutionary conservation of motifs to further filter false positive discoveries (Loots, Locksley et al. 2000, Blanchette and Tompa 2002, Wang and Stormo 2003, Liu, Liu et al. 2004, Sinha, Blanchette et al. 2004, Elemento and Tavazoie 2005, Li and Wong 2005, Li, Zhong et al. 2005). The rationale is that functional motifs should be evolutionary conserved across multiple species (Loots, Locksley et al. 2000), Considering evolutionary conservation as an additional criterion indeed has been shown effective in identifying bona-fide CRE motifs, whereas how to better quantify the evolutionary conservation for CRE motif finding is worth further investigation. For example, the current common practice to quantify the evolutionary conservation of a potential CRE motif is to score the *multiple sequence alignment (MSA)* of its corresponding DNA segments in orthologous sequences. For this quantification strategy, there can be at least two issues. One is that short DNA segments are not always well-aligned with its

corresponding segments in MSA, and consequently, a plethora of not-well-aligned CREs can be missed (Li and Wong 2005). The other issue is that even for DNA-segments that can be aligned well with their corresponding segments, current strategies to score conservation are often debatable. For example, one commonly used strategy is to compare a CRE candidate with its corresponding DNA segments in different orthologous sequences, if the number of mismatches is smaller than a specified cutoff, then the CRE candidate is defined as conserved. Since species divergence time is not considered, strategies like this may result in inaccurate assessment for conservation (Li, Zhong et al. 2005).

We therefore developed the MERCED algorithm to discover conserved CREs between microalgae species *C. reinhardtii* and *V. carteri*. Different from available methods, MERCED defines conserved DNA segments by carefully modeling the species-divergence-time. The algorithm consists of the following steps (see Methods and Materials for details): 1) define orthologous gene pairs in *C. reinhardtii* and *V. carteri* as reciprocal best hits obtained by applying PSI-BLAST to the protein sequences in the two species (Altschul, Madden et al. 1997); 2) construct a nucleotide substitution matrix to model the neutral evolution rate of nucleotide substitution between *C. reinhardtii* and *V. carteri* based on fourfold degenerate sites in proteins of orthologous genes (Li, Wu et al. 1985); 3) define conserved k-mers in regulatory sequences of orthologous genes based on their statistical significance estimated from the nucleotide substitution matrix constructed in step (2). Generating functions are applied here to improve the efficiency for significance calculation(Huang, Kao et al. 2004); 4) group conserved k-mers using hierarchical clustering with average linkage (Sokal and Michener 1985); and 5) define CRE motifs based on patterns of k-mers in the same clusters obtained

in step (4), and define k-mers underlying these patterns as the CRE instances of the corresponding motif. See Figure 2-1 for the pipeline of our method.

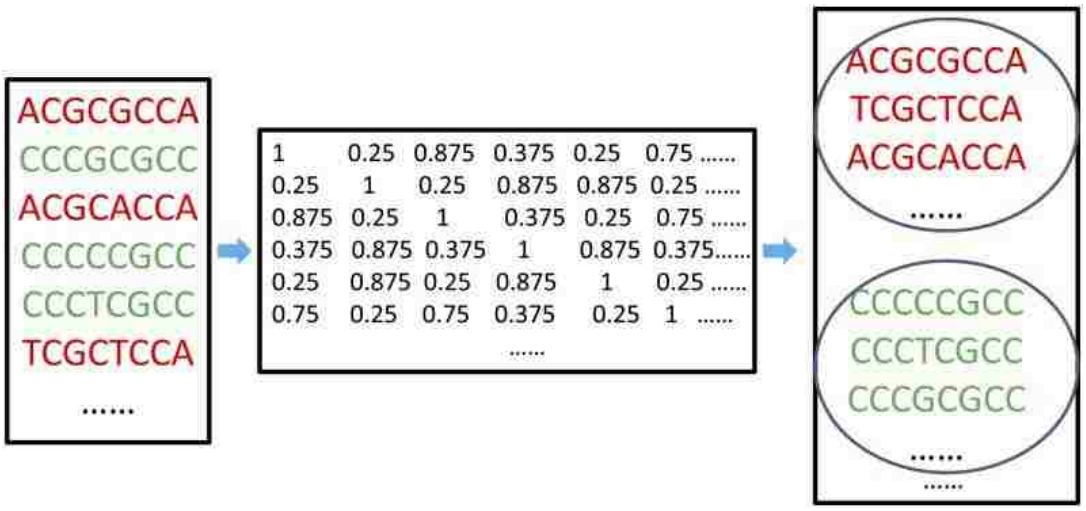

Figure 2-1 Identified CRE motifs from conserved k-mers

The MERCED algorithm has several advantages over MSA-based methods. One of them is that MERCED is more practical for conservation estimation of short DNA segments compared with MSA-based methods. This is because the latter will not work for the not-well-aligned DNA segments, whereas DNA segments, especially the short ones, are often hard to be aligned well using existing MSA algorithms. Additionally, in contrast to MSA-based methods, MERCED considers species-divergent-time rather than merely the number of mismatched nucleotides between DNA segments in aligned orthologous sequences. As a result, we can obtain very accurate estimation of evolutionary conservation even for very short DNA segments. This is because the conservation between two DNA segments from two orthologous sequences depends on not only the number of mismatches, but also the evolutionary distance between them, whereas the absolute value of mismatch number is not always proportional to the evolutionary distance [36]. In fact, highly similar sequences between closely related species, e.g. human and mice, can be present either as a result of active conservation due to functional constraints or as a result of shared ancestry due to insufficient

divergence time (Frazer, Sheehan et al. 2001). For example, as illustrated in Figure 2-2B, one 8-mer in a *C. reinhardtii* sequence with a zero mismatch compared with a similar 8-mer in the orthologous *V. carteri* sequence may have a p-value of 0.011 according to our model and thus will not be considered as a conserved 8-mer pair. On the contrary, another 8-mers with two mismatches compared with its similar 8-mer in the same orthologous sequences will be considered as conserved because the much smaller p-value, 2.84E-05. One additional beneficial side effect of the MERCED is that when comparing potential instances of a CRE motif in different species for CRE conservation quantification, MERCED avoids setting an arbitrary cutoff for the number of mismatches.

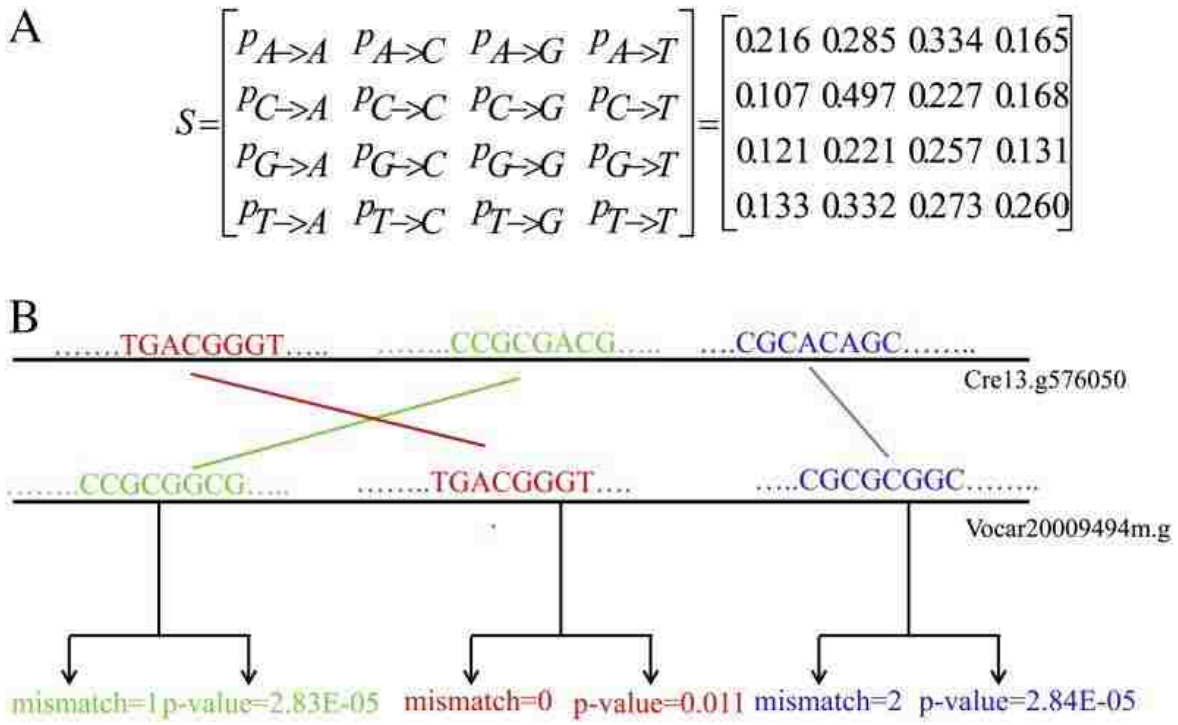

Figure 2-2 A. The substitution matrix that models the neutral evolution rates of nucleotides. B. Three 8-mer pairs as examples to illustrate the smaller conservation P values are not always corresponding to smaller mismatch numbers

2.1.3.2 The predicted motifs are consistent with literature and experimentally verified ones

Applying MERCED to the upstream regulatory sequences of 8741 groups of orthologous genes in *C. reinhardtii* and *V. carteri*, we predicted 317 conserved CRE motifs (False Discovery Rate (FDR) < 0.05 based on permutation) and 66530 corresponding CREs (k = 8). We compared the predicted motifs with the known motifs in the PLACE database (http://www.dna.affrc.go.jp/PLACE/) (Higo, Ugawa et al. 1999) based on the STAMP software tool (Mahony and Benos 2007). PLACE has so far compiled 469 experimentally-verified motifs in plants from literature (Higo, Ugawa et al. 1999), and the STAMP tool is commonly used to assess motif similarities with statistical significance estimation. We found that 195 (62.5%) of the 317 predicted motifs are similar to the PLACE motifs with the STAMP E-value cutoff as 1E-5, which was used in previous studies to define similar motifs (Fauteux, Blanchette et al. 2008, Reed, Charos et al. 2008). Take the predicted motif M75 for example, the consensus of its reverse complement is GNCGGCCA, which is similar to the PLACE motif GRAZMRAB17-GGGCGGCCAGTG (E-value = 4.18E-11). As another example, the predicted motif M78 is similar to the known motif LRENPCABE-ACGTGGCA with the consensus of its reverse complement as CNTGGCA (E-value = 3.5E-11). We also found that 129 (27.5%) PLACE motifs are similar to the predicted 317 motifs and multiple predicted motifs may be similar to the same PLACE motif. For instance, the predicted motifs M4 and M5 have the motif consensus GCAGCTGC and YCAGCAGC, respectively, which are similar to the same known PLACE motif ANAERO2CONSENSUS with the consensus AGCAGC. In the meantime, many of these experimentally verified motifs in PLACE database are not similar to the predicted motifs. One possible explanation is that rather than from the algae, the majority of the PLACE motifs are from land plant species, especially *Arabidopsis* (Higo, Ugawa et al. 1999). Therefore, the difference

between the PLACE motifs and the predicted motifs might suggest the evolved regulatory mechanism during the distant evolution time between the algae and the land plants.

In addition to the comparisons with experimentally verified motifs, we also found literature supports for the predicted motifs. For instance, TGACGCCA is an experimentally verified CRE in the *C. reinhardtii* gene GPX5, which has been shown relevant to the oxidative stress response in *C. reinhardtii* (Fischer, Dayer et al. 2009). This CRE is similar to the consensus of the predicted motif M44, T[CG]C[ACGT]GCCA, where nucleotides included in the parentheses are those frequently occurring at the specified positions. Moreover, we found that the target genes of the motif M44 significantly share the following functions based on gene ontology enrichment analysis (Boyle, Weng et al. 2004): response to stress (GO:0006950, corrected p-value: 7.34E-08), which is consistent with the function of GPX5. Besides, we have also found the predicted motif M44 is similar to PLACE motif GRAZMRAB28 (E-value = 1.78E-08). GRAZMRAB28 is found in the promoters of abscisic acid (ABA) responsive genes (Busk and Pages 1998), which also supports the functionality of M44.

2.1.3.3 The predicted motif combinations are consistent with experimentally verified ones

In high eukaryotes, multiple TFs often coordinately regulate their target genes, by binding to their respective CREs that co-occur in the short regions of a few hundred base pairs (Yuh, Bolouri et al. 1998, Blanchette, Bataille et al. 2006, Ding, Hu et al. 2012). Many co-occurring CREs in plants have also been identified by experimental studies. These CREs can be associated with a pair of interacting TFs or individual TFs with multiple DNA binding domains (Singh 1998). To see whether CREs of the predicted motifs significantly co-occur in the regulatory sequences of *C. reinhardtii*, we applied our previously developed method (Cai, Hou et al. 2010) to identify motif combinations containing

multiple motifs whose instances frequently co-occur in the regulatory sequences. In total, we predicted 92694 motif combinations based on the identified CRE motifs (k = 8), each of which consists of 2 to 6 motifs. The percentage of motif combinations comprised of 2 to 6 motifs is 0.08%, 40.40%, 47.44%, 11.75%, and 0.33%, respectively. As a way to evaluate our prediction, we compared the predicted motif combinations with seven well-known co-occurring motif pairs (Table 2.1) (Singh 1998, Steffens, Galuschka et al. 2005). We found five out of the seven known motif combinations were subsets of our predicted combinations. In general, a known combination can be a subset of multiple predicted combinations. For instance, one of the known motif combinations is comprised of two motifs CAACA and CACCTG that can be bound by the TF RAV1 at the AP2-like N-terminal domain and the B3-like C-terminal domain, respectively. 378 predicted motif combinations have been found to include these two motifs. With different additional motifs, these different motif combinations containing the same known combination can regulate different sets of target genes. For instance, two predicted motif combinations, MOD3428 and MOD3431, both include the above RAV1-binding motif combination of M164 and M27 (Table 2.1). MOD3428 includes one additional motif M11 and MOD3431 includes one additional motif M21. It turns out MOD3428 regulates 204 predicted target genes and MOD3431 regulates 230 predicted target genes, and only 104 of these target genes are shared by both motif combinations, suggesting the additional motifs M11 and M21 play extra roles in these two motif combinations.

Further investigation on the motifs co-occurring with known motif combinations in a predicted motif combination showed that the function of additional motifs relate to that of known motifs in the same combinations. For instance, MOD5159 is a combination of three motifs M13, M167 and M0. The first two motifs comprise the known combination bHLH-MYB and correspond to the binding sites of

TFs with bHLH and MYB protein domains respectively. Both bHLH and MYB-containing TFs have been shown involved in ABA signaling (Abe, Urao et al. 2003). The additional motif M0 is also implicated in ABA signaling because of its similarity to the known PLACE motif AGCBOXNPGLB, which is known to be the binding sequence of stress signal-response factors ERFs (Fujimoto, Ohta et al. 2000). Therefore, the additional motif M0 is also related ABA signaling and its function is consistent with that of the known combination in terms of ABA signaling. In addition, we found that the target genes of this motif combination MOD5159 significantly share the following functional annotation terms: response to stress (GO:0006950, corrected p-value: 1E-4), which further support the coherent function of the three motifs and that of the motif combination.

2.1.3.4 The predicted motif combinations are functionally important

We also performed function study of the predicted motif combinations by investigating the overlap between their target genes and sets of genes with the same GO annotations (see Methods and Materials). Significant overlapping with function-annotated gene sets indicates the function coherence of a motif combination in terms of gene co-regulation (Blanchette, Bataille et al. 2006). In other words, motifs contained in such motif combinations are likely to work together to regulate their target genes. We found that target genes of 87.5% of the predicted motif combinations in *C. reinhardtii* significantly share the same functions (FDR < 0.05) (Materials and Methods). The following two examples illustrated the functional relevance of predicted motif combinations.

Example 1: The motif combination MOD22676 contains four predicted motifs, M39, M4, M25, and M15. All of these four motifs are similar to known motifs that have documented biological-relevance to plant stress response. For example, M4 is similar to the PLACE motif AGCBOXNPGLB-AGCCGCC, which is the binding sequence of stress signal-response factors ERFs (Fujimoto, Ohta et

al. 2000). M39 is similar to the PLACE motif GRAZMRAB28-CATGCCGCC, which was found in the promoter of ABA responsive gene that also play a critical role in response of environmental stress (Busk and Pages 1997). M15 is similar to the PLACE motif ABREMOTIFIIIOSRAB16B-GCCGCGTGGC, which is known to be required for ABA responsiveness as well (Ono, Izawa et al. 1996). Finally, M25 is similar to ABRE3OSRAB1616-GTACGTGGCGC, which is the known CRE motif called ABA responsive element (ABRE) found in rice and related to the biotic and abiotic stresses (Skriver, Olsen et al. 1991). In addition, we also found that the target genes of MOD22676 significantly share the following GO annotation terms: response to stress (GO:0006950, corrected p-value: 4.65E-3), response to oxidative stress (GO:0006979, corrected p-value: 4.85E-3), response to stimulus (GO:0050896, corrected p-value: 0.015), response to chemical stimulus (GO:0042221, corrected p-value: 0.033). These GO functional annotations agree well with the functions of the predicted motifs in this motif combination, supporting the functionality of this motif combination MOD22676.

Example 2: The motif combination MOD9430 is comprised of three predicted motifs, M60, M63 and M22. All the three motifs have similar known motifs that have been reported to involve in photoregulation. For example, M60 is similar to the PLACE motif BOXBPSAS1 that is found in light-regulated asymmetric leaves 1 gene (Fujimoto, Ohta et al. 2000), M63 is similar to the PLACE motif PE3ASPHYA3 with a role in photoregulation (Bruce and Quail 1990), and M22 is similar to the PLACE motif PREMOTIFNPCABE that is also related to photoregulation (Castresana, Garcia-Luque et al. 1988). Additionally, we found that the target genes of MOD9430 significantly share the following GO terms: photosynthesis (GO:0015979, corrected p-value: 0.01). This GO term is

consistent with the functions of the predicted motifs in this motif combination and thus supports the functionality of the predicted motif and motif combination.

In addition to the performed functional study using GO-annotated gene sets, we compared target genes of the 92694 motif combinations with genes that are co-expressed in *C. reinhardtii*. Since co-expressed target genes are often co-regulated (Allocco, Kohane et al. 2004), the significant overlap of target genes of a motif combination with co-expressed gene sets supports the functionality of the predicted motifs and motif combinations as well (Methods and Materials). We found that target genes of 314 predicted motif combinations significantly overlap with at least one co-expressed gene set obtained from four microarray datasets (FDR < 0.05). Consistently, for 289 of the 314 (92.0%) motif combinations, their target genes significantly share the same functions, based on the above GO analysis. In addition, we found that the function of the predicted motifs is likely to be associated with the microarray experimental conditions. For instance, target genes of the motif combination MOD39131 are significantly co-expressed in the gene expression dataset GSE30648, measuring genetic response of *C. reinhardtii* under different oxidative and electrophilic stress conditions. Accordingly, a motif in this combination, M21, is similar to the known motif REGION1OSOSEM, which has been reported to be ABRE (Hattori, Terada et al. 1995) that plays a critical role in environmental stress response (Busk and Pages 1997). Besides, we also found the target genes of MOD39131 significantly share the GO term: response to stress (GO:0006950, corrected p-value=0.016).

2.1.3.5 Comparison with other methods

There is no computational study for genome-wide CRE discovery in *C. reinhardtii*. The MERCED presented here integrates multiple properties of potential CREs as motif-finding criteria. Different

from existing motif-discovery methods applied in other species, the MERCED takes species-divergence-time into account when incorporating sequence-conservation properties to define motifs. To see whether the consideration of species-divergence-time help the CRE motif prediction in *C. reinhardtii*, we compared the MERCED with three alternative approaches using different strategies for sequence conservation evaluation.

The first alternative approach we compared with uses a mismatch number cutoff α to define conserved segments. We implemented this strategy by applying the same procedure of motif-finding as in our method except that we defined conserved k-mers using different mismatch number cutoff α here rather than the generating-function-based statistical significance p-value. In other words, for every pair of promoter sequences corresponding to the 8741 orthologous genes, we defined an k-mer μ in *C. reinhardtii* as conserved if there is one k-mer in μ's corresponding orthologous sequence in *V. carteri* that has at most α mismatches comparing with this *C. reinhardtii* k-mer. We chose the top χ motifs based on the statistical significance, where χ is equal to the number of predicted motifs using MERCED (Liu, Brutlag et al. 2002). We then compared the predicted motifs with known motifs in the PLACE database using STAMP motif-comparison tool (Mahony and Benos 2007) with various E-value cutoffs. We found under each of the STAMP E-value cutoffs, more motifs similar to known motifs are predicted by our method than those predicted by the method based on various mismatch number cutoff α (Table 2.2).

We next compared the MERCED with FastCompare, a more sophisticated approach (Elemento and Tavazoie 2007). FastCompare determines conserved segments directly from sequence comparison, which essentially enumerates all k-mer segments in two species and assesses their significance based

on the difference between their co-occurrence in orthologous sequences of the two species and their occurrence in individual species. We applied FastCompare to our regulatory sequences using the default parameters and only kept the same number of top predicted motifs as predicted using the MERCED for comparison. The results showed our method predicted slightly more motifs that are similar to known plant motifs than FastCompare for various STAMP E-value cutoffs. For instance, for the STAMP E-value cutoff 1E-5, we predicted 195 8-mer motifs that are similar to known motifs in the PLACE database, while FastCompare predicted 169 motifs that are similar to known PLACE motifs. This result indicates the benefit of incorporating species-divergence-time to determine sequence conservation for CRE motif finding in *C. reinhardtii*.

We also compared the MERCED with another alternative strategy used in the PhyloNet (Wang and Stormo 2005). Like FastCompare, PhyloNet is one of the only a few methods that can de novo identify motifs on the genome scale. Briefly speaking, PhyloNet blasts conserved segments in a group of orthologous sequences against conserved segments in all other groups of orthologous sequences to discover motifs. The conserved segments in PhyloNet are defined by the wconsensus algorithm (Hertz and Stormo 1999), which uses the information content to measure the similarity of segments and motifs. After applying PhyloNet to our data and comparing the same number of top ranked motifs predicted by PhyloNet and the MERCED, we found that again, the MERCED predicted more motifs that are similar to known plant motifs than PhyloNet for various STAMP E-value cutoffs (Table 2.2). Not considering the species-divergence-time in PhyloNet can be the partial cause since we observed that many highly similar segments in orthologous sequences that are defined as conserved in PhyloNet may not be conserved if we take species-divergence-time into account. For instance, GAGAAGAA is exactly shared by the regulatory sequences of two orthologous genes,

Cre07.g327250 in *C. reinhardtii* and Vocar2001129m in *V. carteri*. However, it only has a relatively large conservation p-value of 0.057 according to our model. It is thus not considered conserved with the FDR cutoff as 0.05. The overall comparisons shown in Table 2.2 demonstrated the necessity of considering species-divergence-time for CRE motif discovery.

2.1.3.6 A public database for cis-regulatory information in C. reinhardtii

Based on the results obtained in *C. reinhardtii*, we further developed a web-accessible database (http://hulab.ucf.edu/research/projects/Microalgae/sdcre/motifcomb.html), where researchers can download the MERCED software tool and all the currently predicted CRE motifs, CREs, and motif combinations in *C. reinhardtii*. The database supports a variety of queries for specific predicted motifs, motifs similar to PLACE or TRANSFAC motifs, predicted and/or known motif combinations, target genes of specific motifs and/or motif combinations, and so on. A query-based search is able to output the position-weight-matrix (PWM)-represented motifs and their target genes labeled with known expression profiles and GO annotation. The users can choose to view the query results in various formats.

2.1.4 Discussion

Genome-wide identification of CREs in *C. reinhardtii* genome is critical to further study gene regulation and molecular functions. We thus performed the first large-scale CRE prediction in *C. reinhardtii*. Different from available CRE prediction methods, our approach considers the species-divergence-time and the nucleotide content rather than depends on MSA and mismatch-number-counting to determine whether DNA segments are conserved CREs and motifs. According to sequence permutation test, the developed method has very low false discovery rate. Compared with alignment and mismatch-number-counting-based approaches, the developed method is more efficient

in filtering divergent segments while keeping conserved segments that have quite a few mismatches compared with their counterpart segments. In total, we have discovered 66530 CREs corresponding to 317 CRE motifs in *C. reinhardtii* that are conserved in green algae *V. carteri*. We found that 62.5% of the predicted 317 motifs are similar to known PLACE motifs and 27.5% of known PLACE motifs are included in our prediction. In addition, we discovered that the predicted motifs form 92694 statistically significant motif combinations in *C. reinhardtii*. These statistically significant motif combinations are evaluated and supported by known motif combinations, GO enrichment analysis, and gene expression analysis. The large number of CRE motifs and combinations discovered in this study will further facilitate algae research for various applications.

In addition to the genome-wide CRE discovery in *C. reinhardtii*, further analysis of the predicted CREs in *C. reinhardtii* implies that the transcriptional regulation mechanisms may be significantly different between the green algae and the land plant species. We found that two out of the seven well known motif combinations shared by land plant species are not included in our predicted motif combinations. The missing of the well known motif combinations in land plant species suggests that the green algae may have their specific regulatory mechanisms. In fact, whereas genes encoding MADS domains are widespread in land plant genomes [86, 87], there are only two predicted TFs with MADS domains in *C. reinhardtii* (Perez-Rodriguez, Riano-Pachon et al. 2009). This can also be a possible explanation for the lack of motif combinations that contain the well known MADS-MADS motif combinations. Interestingly, the comparisons of the predicted motifs with experimentally-verified known motifs also suggest *C. reinhardtii* may have both plant-like and animal-like motifs. For instance, the predicted motif M3 is similar to several known motifs related to photosynthesis, such as PE3ASPHYA3 (Bruce, Deng et al. 1991). In the mean time, we found that some predicted

motifs tend to have animal-like functions related to flagellum. For instance, The predicted motif M10, not similar to any known PLACE motif, was found in 76 out of 263 (29%) motif combinations whose target genes contain IFT88 (Cre.07.gee5750), which is known to be related to flagellum function (Lucker, Miller et al. 2010). This indicates that M10 may be a novel motif that relates to flagellum function.

Because the precise definition of regulatory sequences is unknown, the regulatory sequences in the current study were restricted to the upstream 1kbp long sequences (Vandepoele, Casneuf et al. 2006, Ma and Bohnert 2007). However, the CREs can be anywhere in the upstream sequences. We compared the results with those obtained from shortened upstream sequences (800bp) and those from extended upstream sequences (1200bp). We made additional predictions from the shortened/extended regions, whereas more than 83.5% predicted motifs are shared for the three different upstream region definitions. Additionally, we used 8-mers motifs to illustrate the MERCED algorithm considering that the most dominant length of motifs in the TRANSFAC database is eight (Wingender, Dietze et al. 1996) and many previous studies have successfully identified meaning motifs in plants and other species using 8-mers (Marino-Ramirez, Spouge et al. 2004, Yamamoto, Yoshioka et al. 2011). We also predicted 7-mer and 9-mer motifs. We found more than 88% of predicted 7-mer and 9-mer motifs are similar to from the discovered 8-mer motifs (STAMP E-value<1E-5 [58]). In the MERCED software, a user can set different length for the regulatory sequences and different parameter *k* for k-mer CRE discovery.

Finally, although we have shown various sources of evidence that support our predictions, these predictions need to be experimentally validated. Currently, the number of genes with annotated

functions in the two algae species is small and experimentally verified CREs are rare. With more and more genomic data available, the prediction accuracy can be further improved and better evaluated. For example, recent sequencing technology has generated more large-scale measurement of gene expression and TF binding and produced ChIP-seq and RNA-seq datasets in many species (Johnson, Mortazavi et al. 2007, Robertson, Hirst et al. 2007). There are already four RNA-seq datasets available in GEO and 15 samples available in Sequence Read Archive in *C. reinhardtii* (Miller, Wu et al. 2010, Castruita, Casero et al. 2011, Fischer, Ledford et al. 2012, Urzica, Adler et al. 2012). Although the number is still small, the data processing strategies e.g., data normalization and isoform identification, are still under development, we can foresee in the near future, a large number of RNA-seq and ChIP-seq datasets can be integrated for CRE prediction and evaluation under different experimental conditions. Additionally, with more genome sequenced, we will be able to integrate more species into our comparative-genomics study, for which, species-divergence-time incorporation is expected to produce more accurate conservation estimation.

2.2 SIOMICS: a Novel Approach for Systematic Identification of Motifs in ChIP-seq Data

2.2.1 Background

Systematic discovery of transcription factor binding sites (TFBSs) and binding motifs is crucial for the study of gene transcriptional regulation (Birney, Stamatoyannopoulos et al. 2007). TFBSs are 6 to 14 base pairs long DNA segments that can be bound by transcription factors (TFs) (Wingender, Dietze et al. 1996, Blanchette, Bataille et al. 2006, Cai, Hou et al. 2010). A TF usually binds to similar TFBSs. The pattern of the TFBSs bound by a TF is called a motif, commonly represented as a position weight matrix (PWM) or a consensus sequence (Stormo 2000). The binding of TFBSs by TFs can activate or repress the transcription of genes near the TFBSs, thus can modulate gene

expression (Arnone and Davidson 1997). In eukaryotes, it is often the TFBSs of multiple TFs in a short DNA region that determines the temporal spatial expression pattern of a gene (Arnone and Davidson 1997). The short DNA regions of several hundred base pairs long that contain TFBSs of multiple TFs are called cis-regulatory modules (CRMs). Correspondingly, we define a motif module as a group of TFs with their TFBSs co-occurring in significantly many CRMs. In other words, a motif module has the TFBSs of all its motifs co-occurring in at least a given number of sequences and has a p-value of motif co-occurring smaller than a given threshold. Because the chance that a short DNA region is a CRM of a motif module is much smaller than the chance that a short DNA segment is a TFBS of a motif, the identification of TFBSs and motifs through the identification of CRMs and motif modules is likely less error-prone than that through the identification of TFBSs of individual TFs (Blanchette, Bataille et al. 2006, Ding, Hu et al. 2012, Ding, Li et al. 2012).

The Chromatin immunoprecipitation followed by massive parallel sequencing (ChIP-seq) experiments provide a great opportunity for computational identification of TFBSs and motifs (Birney, Stamatoyannopoulos et al. 2007, Robertson, Hirst et al. 2007). ChIP-seq experiments can define DNA regions that are enriched with TF binding for a TF under a specific condition on the genome scale. These DNA regions are often called ChIP-seq peak regions. ChIP-seq peak regions, which are averagely several hundred base pairs long, can be identified from ChIP-seq experiments through peak-calling algorithms (Ji, Jiang et al. 2008, Zhang, Liu et al. 2008). Depending on the TF used for the ChIP-seq experiments, there could be several hundred to thousands of peak regions defined in one ChIP-seq experiment. Effective computational methods are necessary to systematically discover motifs and TFBSs of the TF and those of its cofactors. Here and in the

following, a cofactor is a TF that regulates its target genes with the TF used to do the ChIP-seq experiment.

Several computational methods identify motifs in top ChIP-seq peak regions (Jothi, Cuddapah et al. 2008, Valouev, Johnson et al. 2008). Such a type of approaches is likely to miss many potential motifs since TFBSs of cofactors may only occur in some ChIP-seq peaks (Bailey 2011). A few methods attempt to identify TFBSs and motifs in all peak regions, by using known motifs to scan (extended) ChIP-seq peak regions to identify significantly co-occurring motifs (Sun, Guns et al. 2012, Ding, Cai et al. 2013). This type of approaches has achieved certain success in identifying motifs of underrepresented cofactors (Ding, Cai et al. 2013). Since current knowledge of known motifs is still limited, TFBSs and motifs of many cofactors are likely missed by this type of methods. There are also methods for de novo discovery of TFBSs and motifs in all peak regions from a ChIP-seq experiment (Hu, Yu et al. 2010, Kulakovskiy, Boeva et al. 2010, Bailey 2011, Thomas-Chollier, Herrmann et al. 2012). Almost all of this type of methods considers individual motifs separately. Note that TFBSs of certain cofactors may only occur in a small number of peaks (Bailey 2011, Stamatoyannopoulos 2012). Motifs of these cofactors may thus be statistically insignificant individually, as shown in the following analyses. The currently available de novo motif discovery methods may thus miss motifs and TFBSs of many cofactors.

Here we developed a novel computational approach SIOMICS (systematic identification of motifs in Chip-seq data) for de novo discovery of motifs and TFBSs from all peak regions of a ChIP-seq experiment. Instead of considering individual motifs separately for motif discovery, SIOMICS simultaneously considers motif modules, i.e., combinations of any number of motifs that co-occur in

at least a predefined number of peak regions and have p-value of statistical significance smaller than a given threshold. Instead of considering only motifs that are significantly overrepresented in ChIP-seq peak regions, SIOMICS takes both overrepresented and non-overrepresented motifs into account. Tested on 13 ChIP-seq datasets, SIOMICS identified many known motifs, new motifs, and their TFBSs. Tested on 13 simulated random datasets that were obtained by permuting the experimental sequence data, SIOMICS did not predict any false motif. Compared with two recent methods, Dreme (Bailey 2011) and Peak-motifs (Thomas-Chollier, Herrmann et al. 2012) , SIOMICS identified more known cofactor motifs in ChIP-seq datasets and the same or fewer false motifs in random datasets, and had a comparable or better time efficiency.

2.2.2 Materials and Methods

2.2.2.1 ChIP-seq Experimental Data and Simulated Data

We obtained the ChIP-seq experimental data for 13 TFs from Chen et al (Chen, Xu et al. 2008), which were widely used as the benchmark datasets for evaluating TFBS and motif identification methods (Bailey 2011, Thomas-Chollier, Herrmann et al. 2012). We first downloaded the mapped reads from GSE11431 in the Gene Expression Omnibus database (Edgar, Domrachev et al. 2002). We then defined ChIP-seq peaks for each dataset using the peak-calling software MACS (Zhang, Liu et al. 2008). Finally, we obtained the repeat-masked DNA sequences for the defined peak regions using the University of California, Santa Cruz (UCSC) genome browser for each TF (Kent, Sugnet et al. 2002). Note that at this step, to enable TFBSs of more cofactors to be considered, we extended the peaks equally on the two sides of each peak region such that each extended peak region is at least 800 base pairs long. With these experimental sequence datasets, we obtained 13 simulated sequence datasets by randomly permuting nucleotide positions in every obtained sequence in each dataset. In

brief, for a given sequence, say it is n base pairs long, we will randomly generate a permutation of (1, 2, 3, …, n), say (a_1, a_2,…, a_n). We will then move the first nucleotide in this sequence to the a_1-th position of the new sequence, the second nucleotide in this sequence to the a_2-th position of the new sequence, …, the n-th nucleotide to the a_n-th position of the new sequence. In this way, we obtain a new sequence. We repeat this process for every sequence, using an independent permutation each time.

2.2.2.2 Generation of Motif Candidates

SIOMICS identifies motifs by simultaneously considering multiple motifs corresponding to a TF and its cofactors. Since the majority of motifs are still unknown, SIOMICS first obtains motif candidates and then considers the co-occurrence of the motif candidates to define final putative motifs. To generate motif candidates, SIOMICS utilizes k-mers (k base pairs long DNA segments) in the input sequences in a ChIP-seq dataset. Here k=8 was used in the following analyses because an 8-mer can already account for an essential portion of a motif since a common motif is only 6 to 14 base pairs long (Wingender, Dietze et al. 1996). For each k-mer occurring in input sequences, SIOMICS defines it as an k-mer motif candidate by assuming all k-mers in input sequences that are different from this k-mer at most at one position as its TFBSs. SIOMICS then ranks these motif candidates in a ChIP-seq dataset by the following score schema used previously (Liu, Brutlag et al. 2002), from the one with the largest score to the one with the smallest score: guo. Here x_m is the number of TFBSs of a motif candidate, p_{ij} is the frequency of the nucleotide j at position i of the motif candidate and $p_0(s)$ is the probability of generating TFBSs based on background nucleotide frequencies. Other score schemas (Li and Wong 2005, Li, Zhong et al. 2005) have also been tested and do not change the results significantly, which may be due to the fact that final putative motifs in a ChIP-seq dataset are obtained from motif modules. Because many motif candidates may be highly similar to each

other, SIOMICS removes redundant motif candidates with lower ranks such that the consensus sequence of a remaining motif candidate is different from that of other remaining candidates at least at two positions. All remaining motif candidates are used in the following to identify putative motifs.

2.2.2.3 Putative Motif Identification by SIOMICS

With the motif candidates in a ChIP-seq dataset, SIOMICS modifies a frequent pattern mining approach developed in our previous studies (Cai, Hou et al. 2010, Ding, Cai et al. 2013) to discover motifs through the identification of motif modules. The basic idea is to represent motif candidates as nodes in a tree such that more frequent motif candidates are represented at top level (close to the root) and each branch represents the co-occurrence of a group of motif candidates in one or multiple ChIP-seq peaks. Next, an idea similar to the conditional probability is applied to discover groups of co-occurring motif candidates that contain a specific motif candidate and have their TFBSs co-occurring in at least s sequences (Cai, Hou et al. 2010, Ding, Cai et al. 2013). Here is called the support of a group of motif candidates. Finally, a Poisson clumping heuristic strategy (Aldous 1989, Hu, Hu et al. 2008) is implemented to measure the statistical significance of each obtained group of co-occurring motif candidates and output motif modules. The basic idea of this significance calculation is to approximate the occurrence of each motif candidate in sequences by an independent Poisson process and measure how likely we will observe a group of motif candidates occurs in x peak regions, where x>=s (Hu, Hu et al. 2008).For example set s=100, if a group of motif candidates occurs in more than 100 (e.g. 200) peak regions, this group of motif candidates will be kept to be module candidates. Otherwise (occurs only in 90 peak regions), it will be discarded. So a motif module predicted by SIOMICS is a group of motif candidates with their TFBSs co-occurring in at least s peak regions and with the multiple comparison corrected p-value of co-occurrence smaller

than 0.01 or a user specified significance cutoff. The motif candidates in these predicted motif modules are output as the final putative motifs.

Because the large number of peak regions in a ChIP-seq dataset, the number of motif candidates obtained in the above section can be large. Applying the above approach directly to discover motif modules might be time-consuming. To deal with the potential large number of motif candidates above and minimize the time cost, SIOMICS applies the following strategy to discover motifs (Figure 2-3). In brief, with a user specified maximal number of motifs to be identified, say n, firstly, SIOMICS considers the top n motif candidates to discover motif modules. Assume there are n1 distinct motif candidates included in the predicted motif modules. SIOMICS outputs these n1 motif candidates as putative motifs. Next, SIOMICS iteratively identifies other motif candidates that form motif modules with the identified putative motifs, by considering different groups of n motif candidates each time. Each group of n motif candidates always include all putative motifs discovered so far. Finally, if n putative motifs have been predicted or no new putative motifs have been identified after a certain number of iterations, say r iterations, SIOMICS reports all predicted putative motifs, motif modules, and TFBSs, and stop. See the following algorithm for details.

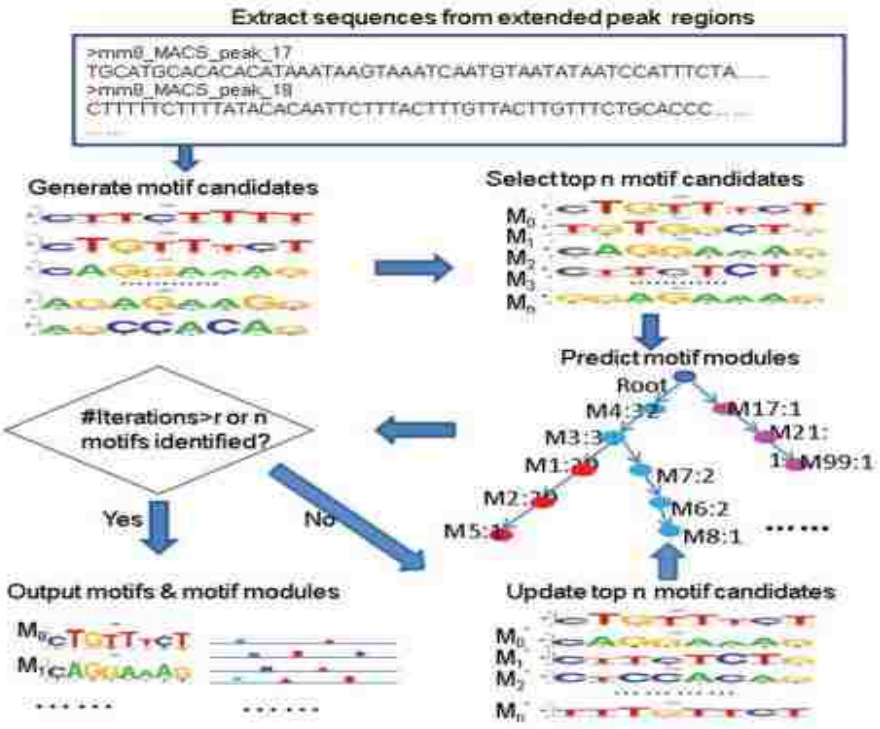

Figure 2-3 SIOMICS Procedure

Algorithm: iterative identification of motifs

INPUT: ChIP-seq sequences, ranked motif candidates, n, r, s

Output: motifs, motif modules, and TFBSs.

Procedure:

1. Initialization: set iteration=0. Discover motifs and motif modules with the top n motif candidates by the above frequent pattern mining approach (Cai, Hou et al. 2010, Ding, Cai et al. 2013), with the support s. Output the n1 motifs included in the predicted motif modules.
2. while n1<n and iteration<r
    3. Discover motifs and motif modules with n motif candidates, which include the n1 motif candidates contained in the predicted motif modules and the top n-n1 motif candidates that have not been used together with the n1 motif candidates for motif module discovery. Output the n2 motifs discovered in the predicted motif modules.
    4. if n2=n1

5. Iteration =0
6. while iteration<r and n2=n1
    7. iteration=iteration+1
    8. Discover motifs and motif modules with n motif candidates, which include the n1 motif candidates and the top n-n1 motif candidates that have not been used together with the n1 motif candidates for motif module discovery. Output the n2 motifs included in the predicted motif modules.
9. n1 = n2.
10. Output the predicted motifs, motif modules, and TFBSs.

2.2.2.4 Cofactors of 13 TF

We obtained known cofactors of the 13 TFs in two ways. One was to extract all cofactors mentioned in (Bailey 2011), which used the ChIP-seq data for the 13 TFs to test the developed motif discovery algorithm. The other was to obtain all interacting TFs for each of the 13 TFs from the BioGRID database (http://thebiogrid.org/) (Chatr-Aryamontri, Breitkreutz et al. 2013), and then confirm each TF by literature search, if they were predicted as a cofactor by any of the 3 software, SIOMICS, Dreme, and Peak-motifs.

2.2.2.5 Comparisons of Predicted Motifs with Known Motifs

For the identified motifs, which were included in certain motif modules, we compared them with known motifs in two public databases, TRANSFAC V11.3 (Wingender, Dietze et al. 1996) and JASPAR 2010 (Sandelin, Alkema et al. 2004). We applied the STAMP tool (Mahony and Benos 2007) with two different E-value cut-offs that were used in previous studies (Ding, Hu et al. 2012, Ding, Li et al. 2012), 1E-4 and 1E-5, for the comparisons of a predicted motif with a known motif.

2.2.3 Results

2.2.3.1 SIOMICS Identifies Known and New Motifs in Each ChIP-seq Dataset

We applied SIOMICS with the default parameters to identify motifs in the 13 ChIP-seq datasets and 13 random datasets. The command used is as follows: "python SIOMICS.py -i seq_fasta -o output_directory -w 8 -m 100 -s 1%*n -r 20 -c 0.01", which means to discover at most 100 motifs of length 8 contained in motif modules that occur in at least 1% of the n peak regions from a ChIP-seq experiment, with the iteration number 20 to stop and the p-value cutoff 0.01 to define motif modules. See the software manual at http://www.cs.ucf.edu/~xiaoman/SIOMICS/SIOMICS.html. SIOMICS identified more than 21 known and new motifs in each ChIP-seq dataset. In addition, SIOMICS predicted no motif in any random dataset, representing a high specificity (Table 2.1).

Table 2.1 Predicted motifs by SIOMICS in 13 ChIP-seq datasets and 13 random datasets

dataset	# peaks	# predicted motifs	# predicted motif modules	%motifs similar to known motifs (Evalue<1E-5)	#motifs similar to known motifs (Evalue<1E-4)	#motifs not in original 100	#motifs predicted in random datasets
Sox2	7761	99	889	78/99=78.8	96/99=97.0	51/99=51.5	0
E2f1	20670	99	2510	79/99=79.8	94/99=94.9	55/99=55.6	0
Stat3	5347	91	1256	72/91=79.1	85/91=93.4	39/91=42.9	0
Nanog	17834	99	1131	76/99=76.8	96/99=97.0	58/99=58.6	0
Oct4	6915	73	719	64/73=87.7	69/73=94.5	42/73=45.2	0
c-Myc	6462	96	1901	74/96=77.1	94/96=97.9	77/96=80.2	0
Klf4	18144	99	2052	83/99=83.8	96/99=97.0	52/99=52.5	0
Ctcf	49114	99	784	78/99=78.8	94/99=94.9	38/99=38.4	0
Zfx	17201	98	1945	75/98=76.5	93/98=94.9	76/98=77.6	0
Tcfcp2l1	45885	71	782	55/71=77.5	68/71=95.8	41/71=57.8	0
Esrrb	49127	43	308	35/43=81.4	41/43=95.3	30/43=69.8	0

n-Myc	10987	94	1766	72/94=76.6	91/94=96.8	80/94=85.1	0
Smad1	2185	21	33	21/21=100	21/21=100	16/21=76.2	0

SIOMICS identified many known motifs in each ChIP-seq dataset. Compared with the known motifs in the TRANSFAC and JASPAR databases (Wingender, Dietze et al. 1996, Sandelin, Alkema et al. 2004), in each dataset, more than 76.0% of the predicted motifs are similar to known motifs, demonstrating that the predicted motifs by SIOMICS are likely to be biologically meaningful instead of arbitrary 8-mer patterns (Table 2.1). On average, in each dataset, more than 62.9% of motifs corresponding to the known cofactors of the TF under consideration are predicted by SIOMICS. Take the Nanog dataset as an example. SIOMICS identified the Nanog motif in this dataset, which occurs in 5.8% of peak regions. In addition, SIOMICS identified six motifs for TFs Sox2, Oct4, Zic3, Klf4, Elf5, and Tead1, all of which are known to cooperate with Nanog to regulate their target genes. Note that the Zic3 and Elf5 TFBSs occur only in 4.5% and 5.2% of peaks, respectively, and are individually not statistically significant enough to be identified if we take the multiple comparisons into account. Since SIOMICS considers multiple motifs simultaneously, it identifies these individually insignificant motifs. For this dataset, SIOMICS identified motifs of seven out of eight known cofactors, demonstrating the success of the systematic discovery motifs in ChIP-seq data by SIOMICS. The only motif missed by SIOMICS is the Essrb motif, which is similar to one of the predicted motifs in this dataset while did not satisfy the required STAMP E-value cut-off (Mahony and Benos 2007) when comparing similarity of the predicted motifs with known motifs.

In addition to motifs corresponding to known cofactors, SIOMICS also identified motifs of potential new cofactors. For instance, SIOMICS identified a motif TTTTAAAA in 3 datasets (Sox2, E2f1 and Nanog). In each dataset, this motif forms a motif module with the same two motifs GAAAGAAA

and CAAAACAA, corresponding to the TFs Hsf (STAMP E-value: 5.3E-06) and Fox (STAMP E-value 2.5E-05), respectively. Hsf has been shown to be interacted with Fox (Amr, Mohamed et al. 2006) and Fox has the function "regulation of RNA splicing" (Fogel, Wexler et al. 2012). Consistently, we found that the target genes of this motif module significantly share the same gene ontology term: regulation of RNA splicing (corrected p-value: 0.0052). Thus, it is likely that the unknown TF corresponding to this new motif may play an important role in regulation of RNA splicing together with Hsf and Fox.

To show that SIOMICS can identify motifs that may be underrepresented in ChIP-seq peak regions, we checked how many percent of predicted motifs were not from the original top 100 motif candidates. As mentioned above, motif candidates were ranked according to their individual statistical significance from the most significant ones to the least significant ones. We found that on average more than 60% of predicted motifs were from motif candidates that were ranked higher than 100, implying that many individually insignificant motifs may play important functional roles (Table 2.1). It also indicates that considering individual motifs separately in motif discovery may miss many functional motifs. For instance, in the Oct4 dataset, SIOMICS identified a motif M67 with motif consensus TCCACCCC, which is insignificant by itself (corrected p-value=1). However, this motif M67 is similar to the motif of the TF Zic2 (NACCACCC, STAMP E-value 1.7E-6), and Zic2 is a known cofactor of Oct4 (Pardo, Lang et al. 2010).

2.2.3.2 SIOMICS Identifies Meaningful Motif Modules in Each ChIP-seq Dataset

SIOMICS discovered a large number of motif modules in each ChIP-seq dataset and no motif module in any random dataset (Table 2.1). The number of motifs is from 2 to 4 in a motif module, with the average of 2.15 motifs per motif module. We investigated the functions of the predicted

motif modules and found that at least 51.5% (65.2% on average) of motif modules in a dataset is partially supported by at least one source of functional evidence.

First, we focused on the predicted motifs that are similar to known motifs to see whether TFs corresponding to their similar known motifs interact.

Table 2.2 Predicted motif modules are supported

dataset	Motif modules contain at least a pair of interacting TF pairs from BioGRID	p-value of Enrichment of TF pairs from BioGRID	Shared motif modules across datasets	Motif modules with preferred motif order (corrected p-value<0.05)	Motif modules supported by at least one type of evidence
Sox2	343/889=38.6%	0	261/889=29.4%	208/889=23.4%	582/889=65.6%
E2f1	1373/2510=54.7%	0	408/2510=16.3%	1452/2510=57.8%	2039/2510=81.2%
Stat3	469/1256=37.3%	0	289/1256=23.0%	244/1256=19.4%	755/1256=60%
Nanog	348/1131=30.8%	0	273/1131=24.13%	428/1131=37.8%	712/1131=62.3%
Oct4	254/719=35.3%	2.2E-271	110/719=15.3%	179/719=24.9%	406/719=56.6%
c-Myc	715/1901=37.6%	0	331/1901=17.4%	506/1901=26.6%	1166/1901=61.3%
Klf4	955/2052=46.5%	0	357/2052=17.4%	1044/2052=50.8%	1517/2052=73.4%
Ctcf	299/784=38.2%	0	181/784=23.1%	402/784=51.3%	584/784=74.5%
Zfx	535/1945=27.5%	0	321/1945=16.5%	762/1945=39.2%	1207/1945=62.1%
Tcfcp2l1	169/782=21.6%	8.8E-136	154/782=19.7%	345/782=44.1%	495/782=63.3%
Esrrb	105/308=34.1%	3.2E-106	51/308=16.6%	125/308=40.6%	204/308=66.2%
n-Myc	807/1766=45.7%	0	311/1766=17.6%	723/1766=40.1%	1249/1766=70.1%
Smad1	11/33=33.3%	4.8E-12	9/33=27.3%	3/33=9.1%	17/33=51.5%

We collected all known TF interactions from the BioGRID database (Chatr-Aryamontri, Breitkreutz et al. 2013) and obtained 648491 interacting TF pairs. For each dataset, we then examined whether motifs of these interacting TF pairs are significantly enriched in the predicted motif modules. We found that the corrected enrichment p-value of known interacting TF pairs is smaller than 1E-10 in all 13 datasets (Table 2.2, columns 2 and 3), demonstrating that SIOMICS predicted many known

interacting TFs in the predicted motif modules.Next, we investigated whether a motif module was predicted in multiple datasets. Since the majority of peak regions in the 13 datasets do not overlap with each other, the repeated prediction of a motif module in different datasets implies the functionality of this motif module. For each dataset, we found a large number of predicted motif modules were shared in at least two datasets (Table 2.2). We provided an example of a motif module consisting of an unknown motif together with those of the interacting TFs Hsf and Fox above. Here is another example. The motif module composed of three motifs, CCTTCCTG, CAAAACAA and CTGCTGGG, were found in the Stat3 and E2f1 datasets, which are similar to the Stat3 motif (STAMP E-value 4.7E-8), the Sox2 motif (STAMP E-value 4.2E-6), and the Ctcf motif (STAMP E-value 2.2E-4), respectively. The interaction between Stat3 and Sox2 was reported previously (Foshay and Gallicano 2008). Sox2 was also shown to be co-working with Ctcf (Donohoe, Silva et al. 2009). In addition to the interactions of Stat3, Sox2, and Ctcf, the three TFs also share similar functions. For instance, Stat3 has the function related to system development (Nakashima 1999). So does Sox2 (Que, Okubo et al. 2007). By analyzing the target genes of this motif module, we found that the target genes significantly share the function "system development" (multiple comparison corrected p-value: 0.03). The functions of the TFs in this motif module are thus consistent with the function of its target genes. All these observations on the TF interactions, the TF functional similarity, and the function consistency of the TFs and the target genes support the functionality of this motif module.

Finally, we examined the relative order of the TFBSs of a pair of motifs in every predicted motif module. The rationale is that if a pair of motifs has their TFBSs in certain preferred order in peak regions, TFs corresponding to this pair of motifs likely interact and the motif module may thus be biologically meaningful. Similar to our previous study (Cai, Hou et al. 2010), for a given motif pairs

in a motif module, we counted in how many peaks the preferred order occurs and then assessed the significance by a binomial test. We found that indeed at least 9.1% of motif modules (or 35.8% on average) have TFBSs of at least a pair of motifs with preferred order of occurrence in ChIP-seq peak regions in each dataset, after multiple comparison correction to define the preferred motif orders (Table 2.2). For instance, in the aforementioned example about the unknown motif TTTTAAAA together with those of the interacting TFs Hsf and Fox, we find that TFBSs of Hsf prefer to bind to the downstream of TFBSs of the Fox motif (Corrected p-value 2.47E-12). The two TFs have been shown to interact (Amr, Mohamed et al. 2006).

2.2.3.3 Comparison with DREME and Peak-Motifs

We compared SIOMICS with DREME and Peak-motifs on the 13 ChIP-seq datasets and 13 random datasets. We used the above default parameters for SIOMICS to output at most 100 motifs.

Table 2.3 Comparison of three methods on prediction of known cofactor motifs

TF	Known motifs found (primary and cofactors) E-value cutoff E-4		
	SIOMICS	DREME	Peak-motifs
Sox2	8/9 (Sox2,Klf4, Stat3, Zic3, Hoxa5, Tcf3, Tead1,Oct4)	8/9 (Sox2, Oct4, Klf4, Stat3,Esrrb, Zic3, Tcf3, Tead1)	4/9 (Sox2,Oct4, Klf4, Esrrb)
E2f1	7/10(E2f1,Stat3, Klf4, Fox, Sp1, Nfkb1, Tbp)	6/10 (E2f1,Stat3, Myc, Klf4, Creb, Sp1)	3/10 (Klf4, Creb, Sp1)
Stat3	6/8 (Stat3,Klf4, Sox2, Myc, Sp1, Irf)	6/8 (Stat3,Klf4, Esrrb, Sox2, Myc,Sp1)	6/8 (Stat3,Klf4, Sox2, Esrrb, Myc, Sp1)
Nanog	7/8 (Nanog,Sox2,Oct4, Zic3, Klf4, Elf5, Tead1)	4/8 (Nanog,Sox2, Klf4, Esrrb)	4/8 (Sox2, Oct4, Klf4, Esrrb)
Oct4	8/10 (Oct4,Sox2, Klf4, Sox10, Ewsr1, Nanog, Zic2, Esrrb)	7/10 (Oct4,Sox2, Klf4, Esrrb, Sox10, Ewsr1, Nanog)	5/10(Oct4,Klf4,Creb, Esrrb, Sox10)
c-Myc	3/4(Stat3, Egr1, Sp1)	3/4 (c-Myc,Stat3, Sp1)	3/4 (c-Myc,Egr1, Sp1)
Klf4	4/10 (Klf4,Stat3, Sox2, Sp1)	6/10 (Klf4,Stat3,Esrrb, Sox2, Sp1, Myc)	3/10 (Klf4,Stat3, Sp1)
Ctcf	5/6 (Ctcf,Stat3,Gabpa, Yy1, Smad3)	4/6 (Ctcf,Stat3,Gabpa, Smad3)	2/6 (Ctcf,Myc)
Zfx	2/4 (Zfx,Stat3)	2/4 (Zfx,Stat3)	2/4 (Zfx,Stat3)
Tcfcp2l1	7/12(Tcfcp2l1,Stat3,Klf4, sox2, Esrrb, Fox, Sp1)	6/12 (Tcfcp2l1,Stat3, Klf4, Esrrb, Fox, Sp1)	5/12 (Klf4, Esrrb, Egr1, Fox, Sp1)
Esrrb	4/10(Esrrb,Klf4, Rxra, Sp1)	8/10(Esrrb,Klf4, Sox2, Stat3, Myc, Rxra, Ewsr1, Sp1)	5/10(Esrrb,Klf4, Stat3, Rxra, Sp1)
n-Myc	2/5 (Stat3,Creb)	2/5(n-Myc,Stat3)	1/5 (n-Myc)
Smad1	5/9(Sox2, Oct4, Esrrb, Klf4, Stat3)	4/9(Sox2, Esrrb, Klf4, Stat3)	4/9(Sox2,Esrrb, Zic3, Klf4)

For Dreme and Peak-motifs, we used the following commands to output at most 100 motifs as well: python dreme.py -p <input_seq> -m 100 -o <output directory>; peaks-motifs -i <input_seq> -prefix

peak_motifs -nmotifs 100 -outdir <output directory>. SIOMICS showed advantages over the two methods in terms of speed and the number of predicted motifs in experimental and random datasets.

We first compared the sensitivity of SIOMICS with that of Dreme and Peak-motifs based on known cofactors of each TF (Table 2.3). In 11 out of 13 the ChIP-seq datasets, SIOMICS did better than or at least the same as Dreme. Only in the Klf4 and Esrrb datasets, Dreme predicted motifs of more known cofactors. Similarly, in 12 out of the 13 ChIP-seq datasets, SIOMICS did at least the same as Peak-motifs. Only in the Esrrb dataset, Peak-motifs predicted motifs of more known cofactors.

To see whether SIOMICS can identify motifs of more cofactors in the Klf4 and Esrrb datasets, we applied SIOMICS to predict motif modules that occur in at least 0.5% of the peak regions instead of the default 1% of the peak regions, SIOMICS identified 2 and 3 motifs of more cofactors in the Klf4 and Esrrb datasets, respectively. For instance, SIOMICS did not identify Stat3, Sox2, and Ewsr1 in the Esrrb dataset at the default 1% cutoff, while identified these motifs when the cutoff 0.5% was used.

We next compared SIOMICS with Dreme and Peak-motifs based on shared motifs predicted by the three methods. This is because we currently have limited knowledge of cofactors of a TF and thus the above comparison of known cofactors may be limited. In addition, if a motif is predicted by at least two of the three independent methods, this motif may be a true motif. To determine whether two predicted motifs by two methods are similar, we required their STAMP comparison E-value be smaller than 1E-5, a more stringent cut-off used in previous studies (Ding, Hu et al. 2012, Ding, Li et al. 2012). We found that for every dataset, SIOMICS predicted much more shared motifs than both

Dreme and Peak-motifs (Table 2.4). Since Dreme and Peak-motifs discover one motif at one time, this comparison implies the advantage of considering multiple motifs simultaneously instead of individual motifs separately.

Table 2.4 Comparison of the three methods on shared motifs

Datasets	SIOMICS	Dreme	Peak-motifs
Sox2	89/99=0.90	60/84=0.71	31/64=0.48
E2f1	92/99=0.93	64/100=0.64	20/49=0.41
Stat3	78/91=0.86	56/70=0.80	27/63=0.43
Nanog	82/99=0.83	60/100=0.60	25/62=0.40
Oct4	63/73=0.86	66/92=0.72	26/57=0.46
c-Myc	84/96=0.88	48/86=0.56	22/59=0.37
Klf4	93/99=0.94	64/100=0.64	27/57=0.47
Ctcf	89/99=0.90	66/100=0.66	20/48=0.42
Zfx	83/98=0.85	58/100=0.58	14/49=0.29
Tcfcp2l1	61/71=0.86	62/100=0.62	17/56=0.30
Esrrb	28/43=0.65	47/100=0.47	15/49=0.31
n-Myc	93/94=0.99	60/100=0.60	17/51=0.33
Smad1	15/21=0.71	19/35=0.54	13/56=0.23

We then compared the specificity of the three methods on 13 random datasets. Since these random datasets were obtained by permuting ChIP-seq peak sequences, they represent sequences with no biological meaning and thus are expected to contain no motif. Indeed, SIOMICS and Dreme predicted no motif in any of these datasets. The fact that no motif was predicted by SIOMICS indicates the small false positive rate can be achieved by simultaneously considering multiple motifs. Although Dreme considers individual motifs separately, it compares the occurrence of a pattern in a ChIP-seq sequence dataset and that in the corresponding permuted dataset (Bailey 2011), which also reduces the false positive rate here. We also found that, on average, Peak-motifs identified 8.62 motifs in a random dataset. We observed that the five datasets with the largest sizes have the larger number of predicted positives by Peak-motifs, which at least partially suggests better false positive control strategies in large datasets by SIOMICS and Dreme.

Finally, we compared the speed of the three methods to discover motifs in the 13 ChIP-seq datasets. All the comparisons were done on the same computer with the following configuration: Intel ® Core™ 2 Duo CPU E7500 @ 2.93GHz and 4G RAM. We found that Peak-motifs is about 1.43 times faster than SIOMICS, which is 15 times faster than Dreme (median). In addition, when the dataset size is small, such as several thousand sequences, the difference between the speed of SIOMICS and that of Peak-motifs is large (around 3 times); when the dataset size is large, the difference between the speeds of the two methods is small (around 1). On the contrary, when the dataset size is large, the difference of the speed of SIOMICS and that of Dreme is large (more than 15 times); when the dataset size is small, the speed difference of SIOMICS and Dreme is small (around 5 times for 5347 peaks). These observations demonstrate the efficiency of SIOMICS in dealing with large datasets (Figure 2-4). It is also implies that when the number of peaks in a ChIP-seq experiment is large, SIOMICS will not only predict motifs of more cofactors than the other two methods, but also have the time efficiency advantage compared with the two methods.

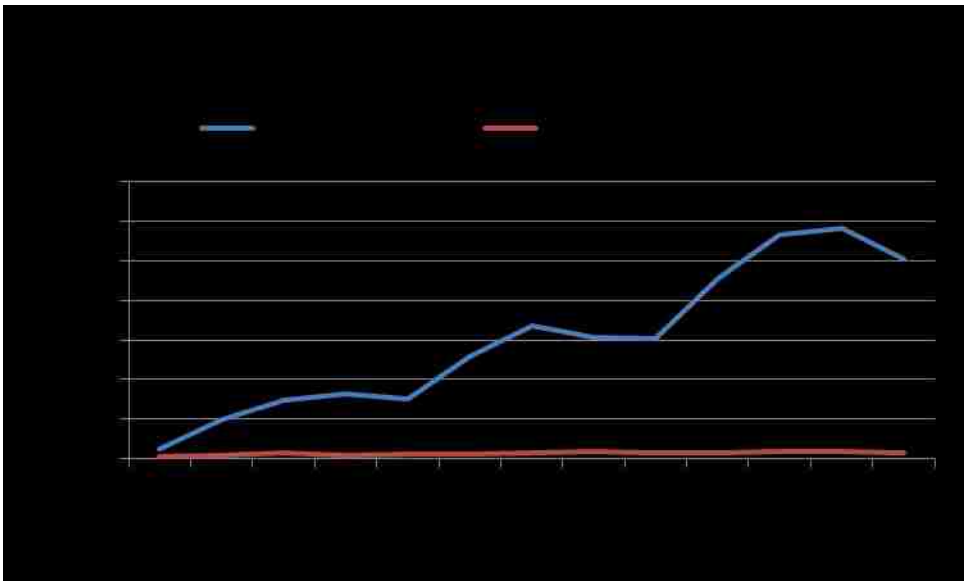

Figure 2-4 Time efficiency comparison

2.2.4 Discussion

We developed a novel approach SIOMICS to systematically discover motifs and TFBSs from ChIP-seq data. Different from available methods, SIOMICS does not depend on limited information of known motifs and simultaneously considers multiple motifs. Tested on experimental and simulated data, we shown that SIOMICS identifies motifs of more known cofactors and identifies more shared motifs in the experimental data. At the same time, SIOMICS has a low false positive rate when tested in the simulated data. In addition, we shown SIOMICS is as fast as other methods especially when the ChIP-seq datasets are large. SIOMICS is thus a useful alternative method for motif discovery.

We applied SIOMICS on the extended 800 base pairs long sequence around the central ChIP-seq peak regions. This is because the central peak regions may not always contain the TFBSs of a cofactor. For instance, for the E2f1 dataset, if we only considered the central peak regions defined by the MACS software (Zhang, Liu et al. 2008), we could have missed the motif of the E2f1 cofactor, Tbp. In the extended E2f1 ChIP-seq peak sequences, SIOMICS identified Tbp as the cofactor of E2f1 (STAMP E-value 1.77E-07). A critical question is how long we should extend the peak regions. Our experience suggests extension of the central peaks such that each peak is at least 800 base pairs long is a good choice. In fact, it has been shown that the majority CRMs are shorter than 800 base pairs (Blanchette, Bataille et al. 2006, Cai, Hou et al. 2010).

In addition to the comparisons with DREME and Peak-motifs, we also compared SIOMICS with CPModule (Sun, Guns et al. 2012), a recently developed tool to discover motif modules using known motifs. As we mentioned above, the number of known motifs is still limited, which is the reason that SIOMICS was developed. Even with the same set of known motifs as input, SIOMICS predicted

motifs of more known cofactors and higher percentages of motif modules supported by data from BioGRID in most of the 13 ChIP-seq datasets within a shorter time.

Users can tune several parameters in the SIOMICS software to optimize the results. The first one is the motif length. We recommend use the motif length 8, and also provide users with the choice to specify their preferred motif lengths from 6 to 14 base pairs. The second parameter is the minimum number of peaks a motif module needs to occur, which was called support above. We used the support as 1% of the total number of peaks in a dataset in the above analysis. The smaller the support is, the more motifs and motif modules are predicted. We recommend using the support as 1% for the speed of the tool. If more information about motif interaction is desired, we recommend use 0.5% as the support. The third parameter is the number of motif candidates considered in an iteration of motif module discovery. We used 100 as an arbitrary cut-off. Users can increase this number if around 100 motifs are predicted, as what was shown in several datasets such as the Sox2 and E2f1 datasets (Table 2.1). We kept the cutoff as 100 for the convenience of the comparisons with Dreme and Peak-motifs. We also tried the cutoff 150 for the several datasets and obtained more motifs of known cofactors. For instance, in the Klf4 dataset, we identified the motif of an additional Klf4 cofactor Tp53 (STAMP E-value 1.7E-05) (Zhang, Geiman et al. 2000), which was not discovered with the cutoff of 100.

In Summary, we developed a novel method for de novo systematic discovery of motifs in ChIP-seq data. This method is shown to predict motifs of more known cofactors than available methods and has comparable speed as the fastest method especially on large datasets. The tool implementing the

developed method, SIOMICS, is freely available at http://www.cs.ucf.edu/~xiaoman/SIOMICS/SIOMICS.html.

2.3 SIOMICS 2: Systematic Discovery of Cofactor Motifs from ChIP-seq Data by SIOMICS.

2.3.1 Background

The discovery of transcription factor binding sites (TFBSs) and binding motifs is critical to understand gene transcriptional regulation (Blanchette, Bataille et al. 2006, Birney, Stamatoyannopoulos et al. 2007). TFBSs are 5-15 base-pair long nucleotide combinations of DNA that are bound by regulatory proteins called transcription factors (TFs), which modulate the expression of genes near the TFBSs. A TF often binds similar TFBSs, the common pattern of which is called a motif (Figure 2-3) (Stormo 2000). In eukaryotic DNA sequences, generally multiple TFBSs occur nearby in a short DNA region that determines the expression of genes (Arnone and Davidson 1997). Such a short region containing multiple TFBSs, normally several hundred base-pair long and rarely more than 2000 base-pair long, is called a cis-regulatory module (CRM) (Arnone and Davidson 1997) (Figure 2-3). We define a motif module as a collection of motifs or TFs with their TFBSs co-occurring in a significant many short DNA regions (i.e., CRMs) (Hu, Hu et al. 2008, Ding, Hu et al. 2012, Ding, Li et al. 2012, Ding, Cai et al. 2013) (Figure 2-5). Statistically, the likelihood of a motif module occurring in *N* sequences is much smaller than that of a motif (Hu, Hu et al. 2008). Therefore, the approach of identifying TFBSs and motifs through prediction of motif modules and CRMs is less error-prone than that of detecting TFBSs from individual TFs.

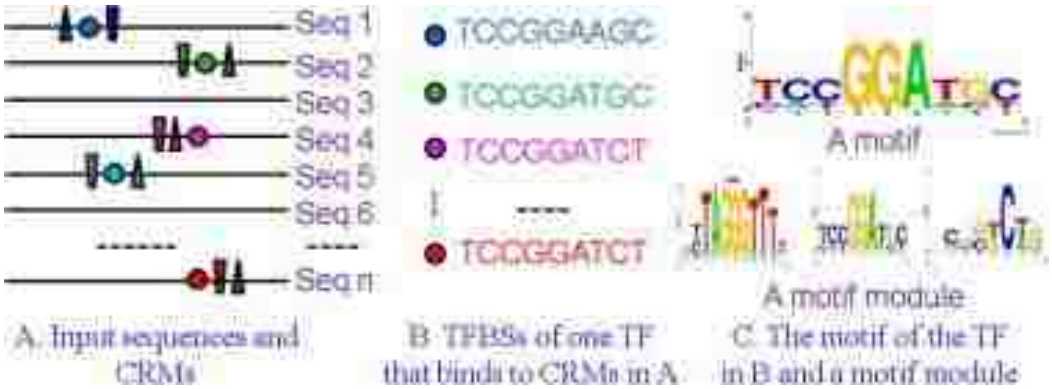

Figure 2-5 An illustration of CRMs, TFBSs, motifs, and motif modules. (A) CRMs are short sequences containing multiple TFBSs. Each line is an input sequence. Each small box on lines represents a TFBS. TFBSs of the same TF are represented with the same shape while with different colors in different sequences. (B) TFBSs of the same TF are similar to each other. (C) The motif corresponding to TFBSs in B and one motif module containing this motif.

Chromatin immunoprecipitation followed by massive parallel sequencing (ChIP-Seq) experiments are generating enormous amount of data that can be harnessed by computational methods to identify biologically-relevant TFBSs and motifs (Johnson, Mortazavi et al. 2007, Robertson, Hirst et al. 2007). With the high-quality antibody to a given TF, ChIP-seq experiments identify TF-binding enriched regions called ChIP-Seq peak regions for this TF by peak calling algorithms such as MACS (Zhang, Liu et al. 2008). ChIP-seq peak regions are in general several hundred base-pair long and enriched with TFBSs of the TF under consideration and its cofactors (Johnson, Mortazavi et al. 2007, Robertson, Hirst et al. 2007). With hundreds or thousands of ChIP-seq peaks, computational methods can then search for motifs and TFBSs in the top ChIP-Seq peak regions, all ChIP-seq peak regions, or neighboring regions around ChIP-seq peaks.

Many computational methods (Jin, Apostolos et al. 2009, Sharov and Ko 2009, Hu, Yu et al. 2010, Kulakovskiy, Boeva et al. 2010, Bailey 2011, Thomas-Chollier, Herrmann et al. 2012, Ding, Cai et al. 2013, Ding, Hu et al. 2013) have been developed to discover TFBSs and motifs from ChIP-seq

data (Table 2.5). These methods can be broadly divided into two types- those considering one motif at a time (Sharov and Ko 2009, Hu, Yu et al. 2010, Kulakovskiy, Boeva et al. 2010, Bailey 2011, Thomas-Chollier, Herrmann et al. 2012) and those considering multiple motifs simultaneously to discover TFBSs and motifs (Jin, Apostolos et al. 2009, Ding, Cai et al. 2013, Ding, Hu et al. 2013). As indicated above, the latter likely predicts motifs and TFBSs more accurately in a ChIP-seq dataset than the former. Here we focus on one of the methods from the second type, SIOMICS (Ding, Hu et al. 2013). To the best of our knowledge, SIOMICS is the first method that considers multiple motifs simultaneously and de novo discovers motifs and TFBSs in a ChIP-seq dataset. It also enhances the prediction of cofactor motifs and under-represented motifs and TFBSs from ChIP-Seq data.

Table 2.5 Commonly used motif prediction tools for ChIP-seq data analysis

Program	Notes	URL
DREME	Discriminative motif finder	http://meme.nbcr.net/meme/doc/dreme.html
HMS	incorporates sequencing depth information	http://www.sph.umich.edu/csg/qin/HMS/
ChIPModule	Known Motif weight-matrices as input	http://www.cs.ucf.edu/~xiaoman/ChIPModule/ChIPModule.html
SIOMICS	De-novo motif discovery	http://www.eecs.ucf.edu/~xiaoman/SIOMICS/SIOMICS.html
CisFinder	Position Frequency Matrix	http://lgsun.grc.nia.nih.gov/CisFinder/
RSAT peak-motifs	Combines four methods	http://rsat.ulb.ac.be/peak-motifs_form.cgi
W-ChIPeaks	Probe-based peak enhancement and known motifs	http://motif.bmi.ohio-state.edu/W-ChIPeaks/
ChIPMunk	Fast heuristic motif digger	http://autosome.ru/smbsm/librettos/libretto_chipmunk/chipmunk_details.rhtml

In the following, we will first describe the original SIOMICS method (SIOMICS) and its extension (SIOMICS_Extension) in the Material and Methods section. Consequently we discuss the performance of SIOMICS_Extension on ChIP-seq datasets in the Results section. Next, in the Results section, we will discuss how to apply SIOMICS_Extension to next generation sequencing

datasets for motif and TFBS discovery. Finally, we provide more insight about SIOMICS_Extension in the Discussion section.

2.3.2 Material and Methods

2.3.2.1 The Original SIOMICS Methods

The methods implemented in the original SIOMICS tool has been delineated in (Ding, Hu et al. 2013). In brief, it starts from a group of input sequences (Figure 2-6). These sequences are from the extended regions centered on the original ChIP-seq peak regions. The reason for extending the ChIP-seq peak regions is that certain cofactor TFBSs may not be included in the original ChIP-seq peak regions (Ding, Hu et al. 2013). With the input sequences, SIOMICS ranks all *w*-mer (*w* base pair long DNA segment, default *w*=8) patterns in input sequences, based on their likelihood ratio scores (Liu, Brutlag et al. 2002) from the largest to the smallest. SIOMICS then iteratively chooses top *m* *w*-mer patterns to count their co-occurrence frequencies (Ding, Cai et al. 2013) and assess the statistical significance of the co-occurrence frequencies (Hu, Hu et al. 2008). The top *m* *w*-mer patterns used in every iteration are different, so that individually insignificant patterns may be identified as motifs because of the co-occurrence of the instances of groups of patterns. Each group of statistically significant co-occurring *w*-mer patterns discovered in an iteration is output as a predicted motif module, with each *w*-mer pattern in the predicted motif modules as a predicted motif. The co-occurring instances of the *w*-mer patterns of a motif module in input sequences are defined as the TFBSs of the motifs in this motif module. The iteration is repeated until *m* motifs are discovered or no new motif is found in *r* consecutive iterations. Here, *m* and *r* are parameters input by users. At the end, there will be no more than *m* predicted motifs, each of which is *w* base pairs long (default *w*=8).

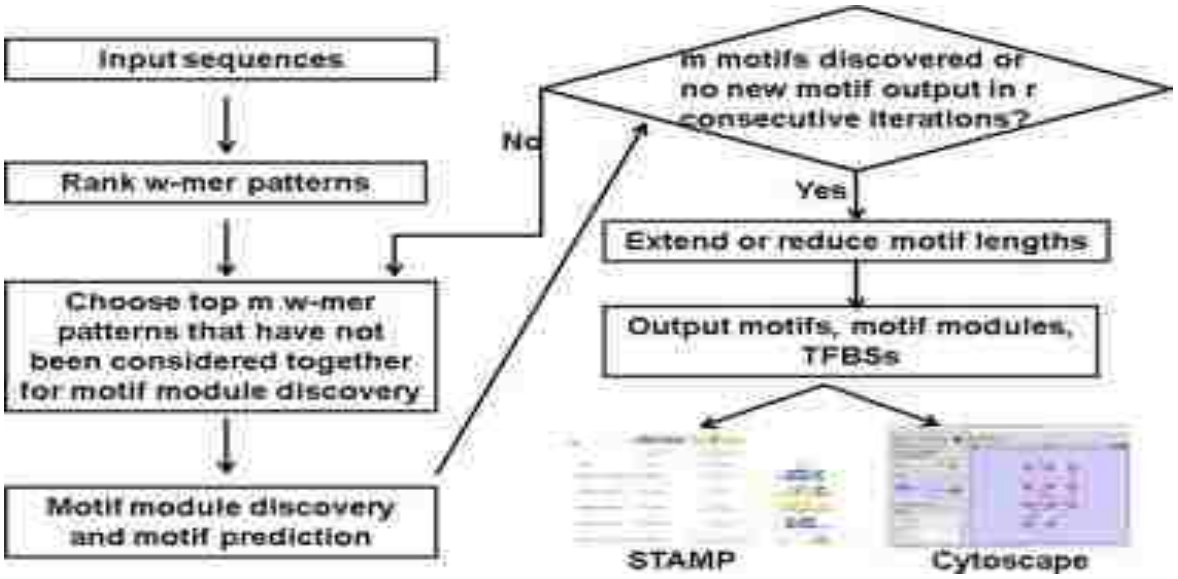

Figure 2-6 The major steps in SIOMICS_Extension

2.3.2.2 The Extended SIOMICS Methods

The major improvement of the SIOMICS tool in the extended version is its ability to output motifs of different lengths. In practice, the common motif length varies from 5 to 15 base pairs (Wingender, Dietze et al. 1996). The initial SIOMICS tools may thus cause certain problems by requiring the predicted motif length be always *w* base pairs long.

To address this issue, SIOMICS_Extension first considers extending these motifs from the left side, or the right side, or both sides. For each discovered motif, SIOMICS_Extension collects all its predicted TFBSs in input sequences. Note that many segments in input sequences that are exactly the same as the predicted TFBSs of this motif may not be considered as the predicted TFBSs, due to the fact that they do not co-occur with the TFBSs of other motifs in the predicted motif modules containing this motif. With the predicted TFBSs of this motif, SIOMICS_Extension then considers the neighboring nucleotides around the predicted TFBSs. The intuition behind this approach is that if this motif is longer than *w*, it is expected that the neighboring nucleotides around most predicted

TFBSs likely share a common nucleotide, on the left side, on the right side, or on both sides. SIOMICS_Extension applies a binomial test to assess whether the neighboring nucleotides are significantly sharing a type of nucleotides and then decides whether to extend the motif length by 1. SIOMICS_Extension extends the motif according to the above principle, one position at a time, from the left side to the right side of the currently predicted TFBSs. When no extension can be made, SIOMICS_Extension output the final form of the motif with the corresponding TFBSs. If no extension is made from the beginning on either side, SIOMICS_Extension also considers extending two adjacent positions at a time and repeats the above process until no more extension can be made.

After extending the motifs, there may be some motifs left unchanged. For these motifs, SIOMICS_Extension considers reducing their motif lengths. SIOMICS_Extension compares these unchanged motifs to see whether two unchanged motifs are similar. Moreover, SIOMICS_Extension checks whether these similar motifs form motif modules with a common group of other motifs. The following are the detailed procedure of Extending and Reducing:

Extending:

Inputs: TFBSs of all predicted motifs

1st Step: Get neighboring nucleotides (8 nucleotides left and right respectively) around the predicted TFBSs of each motif

2nd Step: Extend left by binomial test

Extend rules:

Perform binomial test to get the dominant nucleotides in each position (Position X)
The total # of nucleotides in specific position X=> NT
# of nucleotides (NX) in specific location X=> NX
NX belongs to [A, C, G, T]
The background frequency of A, C, G, T were learned from the genome of this species=> bg (A/C/G/T)
P-value (X) =1-pbinom (NX-1, NT,bg(NX)) (multiple comparisons were performed for these p-values)
If P-value (X)<**pcut:**
->Keep extending position X+1
else
->perform binomial test for the 2 adjacent positions starting with position X (X and X+1)
If there are **NO** dominant 2-mers for those 2 adjacent positions at X
-> **Quit** extending
else
-> **Keep** Extending on position X+2

3rd Step: Extend right by binomial test with the same Extend rules

Output: (1) Extended motifs
(2) Un-extended motifs

Reducing:

Inputs: un-extended motifs

1st Step: Get those unchanged motifs after extending procedure => Set U

2nd Step: Compare every pairs (***a, b***) of motifs from *U*

If these 2 motifs only have one mismatch for their consensus sequences
One mismatch can be the following 3 conditions:
(1) The prefix of ***a*** **(*a*** except for the last nucleotide) was found exactly in ***b***
(e.g. *a*: **ACGTAG**T, *b*: T**ACGTAG**)
(2) The suffix of ***a*** (***a*** except for the first nucleotide) was found exactly in ***b***
(e.g. *a*: T**CGTAGT**, *b*: **CGTAGT**A)
(3) ***a*** and ***b*** have one mismatch
(e.g. *a*: C**A**G**TAGT**, *b*: C**C**G**TAGT**)
If they also have the same co-factors
-> These 2 motifs will be **merged** into one motif.
The PWM of the new motif will be built based on the shared parts of these 2 motifs.
e.g.(1) *a*: **ACGTAG**T, *b*: T**ACGTAG => s**hare: ACGTAG
(2) *a*: T**CGTAGT**, *b*: **CGTAGT**=> share: CGTAGT
(3) *a*: C**A**G**TAGT**, *b*: C**C**G**TAGT**=> share: C?GTAGT (? Depends on the TFBS instances)

Outputs:

Reduced motifs

If both criteria are satisfied, SIOMICS_Extension considers whether to merge these similar motifs and represent them with the similar portion, by a similar binomial test implemented in the extension step.

To better illustrate the above procedures, we provided the following extending and reducing example.

Extending:

For example, assume motif *a* (consensus: CCTGCTGG), the left neighboring position 1of motif *a* is dominant by letter 'C' (binomial test, p-value cutoff: 1e-10, other cutoffs like 1e-12, or 1e-5 were also tried, results are similar). We continue to check the left neighboring position 2 of motif *a*, there is no dominant letter was found at this position and then we tried 2 left adjacent positions starting from left neighboring position 2, there is still no dominant 2-mers. Based on the extension rule, the extending will be ended. Similarly, we do the same extending to the right positions of motif *a*. Finally, we found that we can only extend motif *a* to the left 1 position, the motif *a* will be extended to motif *a'* (consensus: CCCGTCTGG). The PWM of extended motif *a'* is obtained based on the TFBS instances (-1 position -> end of the TFBS instances).

Reducing:

For example, suppose we have identified 2 unchanged motifs after extending, they are:

motif *x* (consensus: ACGGAATG)

motif *y*(consensus: CGGAATGT)

We compare motif *x* and *y*, and we found that they only 1 mismatch and sharing a consensus sequence CGGAATG. Besides, we have also found there are 2 motif modules (x,z) and (y,z) , which means motif x and y have the same co-factor z. Based on the reducing rule, motif x and y will be merged into one motif xy (consensus: CGGAATG), the PWM of motif xy can be obtained from the

TFBS instances of original motif x and y. These 2 motifs (with length of 8) were merged into one motif with length of 7.

After extending and reducing motif lengths, SIOMICS_Extension outputs the final motifs, motif modules, and TFBSs (Figure 2-6). The final motifs are those obtained from the extension and reduction procedure. The motif modules are produced by running the SIOMICS_Extension with the final motifs without any iteration. The TFBSs of the final motif modules are output as the final TFBSs. The final predictions are then visualized using STAMP and Cystoscape tools (Shannon, Markiel et al. 2003, Mahony and Benos 2007).

2.3.3 Results

2.3.3.1 The Performance of SIOMICS_Extension on ChIP-seq Datasets

We previously applied SIOMICS to 13 real ChIP-seq datasets and 13 random datasets (Ding, Hu et al. 2013). We showed that 76.8% to 100% of the predicted motifs by SIOMICS in the 13 real ChIP-seq datasets were similar to known motifs in public databases (Wingender, Dietze et al. 1996, Sandelin, Alkema et al. 2004). The similarity of the predicted motifs to known motifs was measured by using the STAMP software (Mahony and Benos 2007) with the similarity E-value cutoff of 1E-5. Moreover, many individually underrepresented motifs in the 13 real ChIP-seq datasets were discovered by SIOMICS. We also found that no motif was predicted in any of the 13 random datasets of the same sizes. Compared with two recent methods for motif discovery on ChIP-seq data (Bailey 2011, Thomas-Chollier, Herrmann et al. 2012), we showed that SIOMICS predicted motifs of more cofactors with a comparable or better speed in real datasets.

Here we applied the SIOMICS_Extension to the same 13 real ChIP-seq datasets with the default parameters. Since changing motif lengths may make the predicted motifs more similar to true motifs, we paid attention to how many more known cofactor motifs were predicted and the time cost to finish the prediction on each dataset. In Table 2.6, we listed the name of the known cofactors predicted by the original SIOMICS and by the extended SIOMICS_Extension version. The SIOMICS_Extension predicts more cofactor motifs in 5 out of 13 datasets and predicted the correct motifs for 10 more cofactors in these datasets. For instance, for the Esrrb dataset, the original SIOMICS predicted motifs of 3 known cofactors of Esrrb together with the Esrrb motif, while the SIOMICS_Extension predicted motifs of 4 more cofactors in addition to the four motifs predicted by the original SIOMICS tool (Table 2.6).

Table 2.6 Comparison of predicted cofactors by original and SIOMICS_Extension

Data set	SIOMICS	SIOMIC Extension
Sox2	8/9(Sox2,Klf4,Stat3,Zic3,Hoxa5,Tcf3,Tead1,Oct4)	8/9(Sox2,Klf4,Stat3,Tead1,Oct4,Zic3,Tcf3,Hoxa5)
E2f1	7/10(E2f1,Stat3,Klf4,Fox,Sp1,Nfkb1,Tbp)	8/10(E2f1, Stat3,Klf,Fox,Sp1,Nfkb1, Creb1,Tbp)
Stat3	6/8(Stat3,Klf4,Sox2,Myc,Sp1,Irf)	6/8(Stat3,Klf4,Sox2,Myc,Sp1,Irf)
Nanog	7/8(Nanog,Sox2,Oct4,Zic3,Klf4,Elf5,Tead1)	8/8(Nanog,Sox2,Oct4,Zic3,Klf4,Esrrb,Elf5,Tead1)
Oct4	8/10(Oct4,Sox2,Klf4,Sox10,Ewsr1,Nanog,Zic,Esrrb)	8/10(Oct4,Sox2,Klf4,Sox10,Ewsr1,Nanog,Esrrb,Tead1)
c-Myc	3/4(Stat3,Egr1,Sp1)	3/4(Stat3,Egr1,Sp1)
Klf4	4/10(Klf4,Stat3,Sox2,Sp1)	5/10(Klf4,Stat3,Sp1,Myc,Sox2)
Ctcf	5/6(Ctcf,Stat3,Gabpa,Yy1,Smad3)	5/6(Ctcf,Stat3,Gabpa,Smad3,Myc)
Zfx	2/4(Zfx,Stat3)	2/4(Zfx,Stat3)
Tcfcp2l1	7/12(Tcfcp2l1,Stat3,Klf4,Sox2,Esrrb,Fox,Sp1)	10/12(Tcfcp2l1,Stat3,Klf4,Sox2,Fox,Sp1,Oct4,Creb,Myc,Tead1)
Esrrb	4/10(Esrrb,Klf4,Rxra,Sp1)	8/10(Esrrb, Klf4,Rxra,Sp1,Ewsr1,Creb,Sox2,Stat3)
n-Myc	2/5(Stat3,Creb)	2/5(Stat3,Creb)
Smad1	5/9(Sox2,Oct4,Esrrb,Klf4,Stat3)	5/9(Sox2,Oct4,Klf4,Stat3,Esrrb)

We also studied the time cost of the SIOMICS_Extension. All calculations were done on the same computer with the following configuration: Intel Core ™ 2 Duo CPU E7500 @ 2.93GHz and 4GRAM. The running time of the SIOMICS_Extension was about 31.4% more than that of the original version on average, with the median increment of 10.5%. For the largest dataset, the Esrrb dataset, the running time was 51412 CPU seconds. This dataset contains 49127 sequences, each of which is 800 base pairs or longer. For a smaller dataset that contains 5347 sequences, the Stat3 dataset, the running time was 8775 CPU seconds (about 2.44 CPU hours). Note that the time cost of the SIOMICS_Extension was still lower than that of Dreme (Bailey 2011), a popular method for motif discovery on ChIP-seq data, on the corresponding datasets. We also noticed that the increment of the running time becomes smaller when the size of the datasets increases. For instance, for the aforementioned largest dataset, the time cost only increased 4.7%. On the other hand, for the aforementioned smaller dataset, the time cost increased 64.1%. The detailed running speed comparison for SIOMICS, SIOMICS_Extension and Dreme were provided in Figure 2-7.

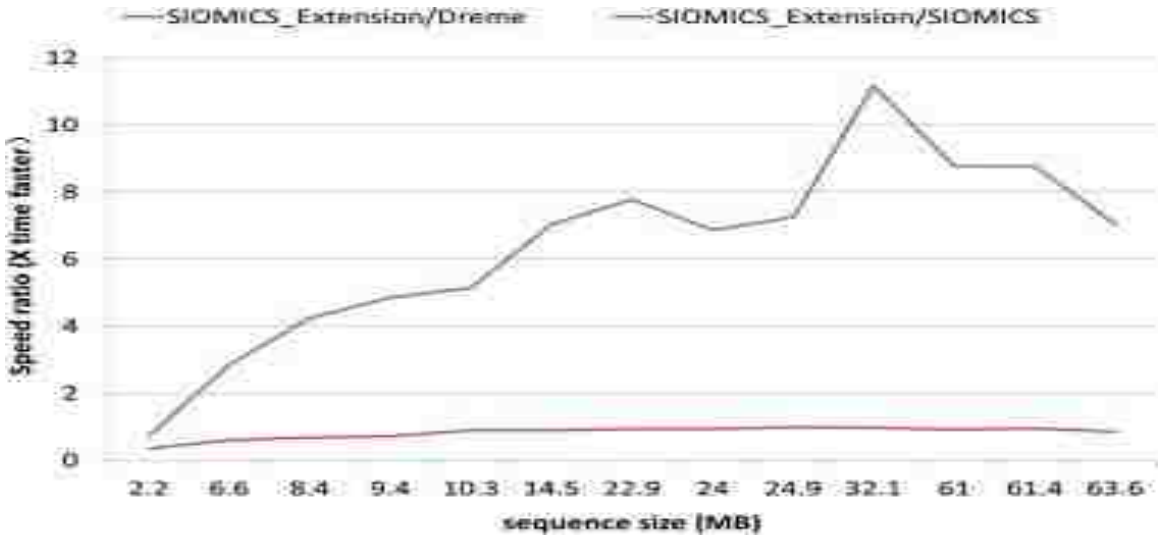

Figure 2-7 Time efficiency comparison of SIOMICS_Extension

2.3.3.2 The Prerequisites to Run SIOMICS_Extension

SIOMICS_Extension is a multiplatform too Time efficiency comparison between SIOMICS_Extension, SIOMICS and Dreme that comes packaged with a graphical user interface (GUI) and a command line interface along with extensions to make comparative analyses. The following requirements must be met in order to successfully setup and run the SIOMICS_Extension tool:

(1) Users need to install the Python 2 or Python 3. If the Linux version is desired, in order to enable GUI, users need to install Tkinter (Python 2) or tkinter (Python 3).

(2) Users need to install Java Runtime Environment.

(3) Users may also need to configure the Environment Variable for running java and python.

The detailed instructions of how to set up the required environment has been provided in the manual of the SIOMIC_Extension tool.

2.3.3.3 The Input to SIOMICS

The input to the SIOMICS_Extension tool is the name of a sequence file together with 7 other parameters. The sequence file must be in the FASTA format, with the sequence length around 800 base pairs. Although SIOMICS_Extension can deal with longer sequences, it is recommended that the majority sequences should be shorter than 1500 base pairs.

Table 2.7 summarizes the major eight parameters in SIOMICS_Extension along with their functions. Among these parameters, *-w* controls the length of motifs before extension and reduction. We recommend use the default *w*=8, based on our experience on motif discovery (Ding, Li et al. 2012, Ding, Hu et al. 2013). The parameter *–m* controls the maximal number of motifs predicted. It also specifies the number of top *w*-mer patterns considered for motif module discovery in each iteration.

Users can start to run SIOMICS with *m*=100. If close to 100 motifs are discovered, say 95, users can increase *m* by 50 and try to run SIOMICS again. We rarely see more than 200 motifs predicted in a ChIP-seq dataset using an antibody against a TF. The *–c* parameter specifies the multiple-comparison-corrected p-value cutoff. The *–s* parameter tells how frequent the co-occurrence of *w*-mer patterns needs to be. The smaller *–s* parameter is, the longer the running time SIOMICS takes. We normally expect that each unknown cofactor bind at least 1% of the ChIP-seq peak regions and thus set the default *–s* parameter correspondingly. The *–r* parameter does not affect the results much. With larger *–r* parameter, users can obtain more comprehensive predictions with higher time cost. The default *–r* parameter is also recommended be used.

Table 2.7 Initial parameters for SIOMICS

Command line argument	Function
-i	Input sequence file
-o	The output folder for prediction results
-w	Initial motif length to search
-m	Maximum number of motifs to output
-s	Minimum number of sequences needed to contain co-occurring TFBSs of a motif module
-c	Multiple comparison corrected P-value cutoff for motif module prediction
-r	Maximum number of iterations with no new motif discovered allowed

2.3.3.4 The Output of SIOMICS_Extension

SIOMICS_Extension output seven files, each of which is briefly described in Table 2.8. The predicted motifs are included in the file named X.motifs, which provides the motif position weight matrix together with other information that can be used to scan sequences to define putative TFBSs. If users want to know which known motifs are similar to these predicted motifs, they can examine

two other files named X.JASPAR.pdf and X.TRANSFAC.pdf. These two files include the comparison results of the predicted motifs with the known motifs in the JASPAR (Sandelin, Alkema et al. 2004) and TRANSFAC databases (Wingender, Dietze et al. 1996). The file X.mc contains all predicted motif modules. For each motif module, this file provides the motifs included in this module, the number of sequences containing the co-occurring TFBSs of all motifs in this module, and the statistical significance of this module represented as the multiple-comparison-corrected p-value. If users want to explore the global picture of the motif (TF) interaction under the current experimental condition, they can utilize Cytoscape (Shannon, Markiel et al. 2003) to open the file named X.mc.sif. This file presents the predicted motif modules in a network, in which motifs (TFs) are denoted as nodes and two motifs are connected if they co-occur in at least one predicted motif module. The seventh file is the running log file, which stores the running time and the command used for a dataset under consideration.

Table 2.8 Output files from SIOMICS

Output files	Description
X.motifs	Predicted motifs from SIOMICS
X.mc	Predicted motif modules
X.JASPAR.pdf	predicted motifs compared with JASPR motifs
X.TRANSFAC.pdf	predicted motifs compared with TRANSFAC motifs
X.mc.sif	Interaction network to be viewed in Cytoscape
X.tfbs	TFBSs of the predicted motifs
running.log	Runtime and command log of SIOMICS

2.3.3.5 SIOMICS_Extension Workflow

Here we provide a brief walk-through for a normal use case of the SIOMICS_Extension application. More details can be found on the software manual page.

1. The data to be processed needs to be in FASTA format, with each sequence in the following format:

   > Oct4 <Sequence identifier>

   ACGTACT…… <sequence nucleotide information>

   See the software manual page for detailed examples. This FASTA sequence file is recommended be repeat masked at http://www.repeatmasker.org/ or locally.

2. With the sequences, we can open the SIOMICS with the default parameters to start the motif and TFBS prediction. See the interface opened on the software manual page or at http://www.eecs.ucf.edu/~xiaoman/SIOMICS/man_Extension.html.

3. The GUI window near the bottom will display a message, "Running…" to indicate that the software has begun motif discovery. It will display "Done" with all the output files and the corrections made.

4. If Cytoscape is installed and the extensions are enabled, the end of processing in **SIOMICS_Extension** will take users directly to the interaction network mapping. Otherwise, users can open the X.mc.sif file using the Cytoscape software (Shannon, Markiel et al. 2003).

2.3.4. Discussion

We developed a useful tool set SIOMICS for systematic discovery of motifs in ChIP-seq datasets. Tested on 13 real ChIP-seq datasets and 13 random datasets, we showed that SIOMICS predicted motifs of known cofactors in each real dataset while no motif in any random dataset (Ding, Hu et al. 2013). We also compared SIOMICS with two motif discovery methods developed for ChIP-seq data analysis (Bailey 2011, Thomas-Chollier, Herrmann et al. 2012). SIOMICS Showed superior or at least comparable performance in terms of the number of motifs of known cofactors predicted in real

datasets, false motifs predicted in the random datasets, and the time cost of the predictions (Ding, Hu et al. 2013). With GUI, SIOMICS will be a useful tool for the systematic study of gene transcription regulation based on ChIP-seq experiments.

In the extension version of the SIOMICS tool –SIOMICS_Extension, we modified the motif length of a motif based on its predicted TFBSs from all motif modules containing this motif. Alternatively, we can implement similar procedures using all its TFBSs from individual motif modules containing this motif. In this way, we might identify different forms of the same motif when this motif co-occurs with different cofactor motifs. However, given the large number of the predicted motif modules, this process may take a large amount of time. Moreover, there may be not many motifs with different forms when they cooperate with different cofactors. In fact, tested on the 13 real datasets, we rarely found different forms produced from the same *w*-mer motifs. We thus consider all TFBSs of a motif from all motif modules containing this motif.

Although SIOMICS tool set (Original SIOMICS and SIOMICS_Extension) is developed for ChIP-seq data, it also works well on small datasets not from ChIP-seq experiments. We tested SIOMICS on small datasets with around a dozen to several dozen 1000 base-pair long sequences (Chang, Sneddon et al. 2004, Wang, Klijn et al. 2005, Weigelt, Hu et al. 2005, Li, Cheng et al. 2006). SIOMICS identified motifs of several TFs shared by these independent breast cancer related datasets. Note that when the number of sequences is small, we recommend that a relative larger *–s* parameter should be used to control the false positive rate of the predictions.

In summary, we developed a useful tool for gene transcriptional regulation studies. The developed SIOMICS tool set (Original SIOMICS and SIOMICS_Extension) together with its manual, test datasets, and others is freely available at http://www.cs.ucf.edu/~xiaoman/SIOMICS/SIOMICS.html. We will continue to keep the updated versions on this site to help researchers fully utilize the potential of the ChIP-seq data.

CHAPTER 3: POST TRANSCRIPTIONAL REGULATON-MICRORNA BINDING

3.1 miRModule: MicroRNA Modules Prefer to Bind Weak and Unconventional Target Sites

3.1.1 Background

MicroRNAs (miRNAs) play critical roles in gene regulation (Bartel 2004, Bartel 2009). MiRNAs are a family of small (~22 nucleotides) non-coding RNAs. They can bind mRNAs at 5' untranslated regions (UTRs), coding sequences (CDSs), and 3' UTRs. The binding is traditionally thought to be through base-pairing of the seed regions in miRNAs with the partially complementary sequences in target mRNAs (Bartel 2009). The seed region refers to the 5' end of miRNAs from position 2 to position 7 (Lewis, Shih et al. 2003, Lewis, Burge et al. 2005). Depending on the pairing quality, miRNA target sites are classified into two categories: canonical sites and non-canonical sites. The former are target sites that are perfect complementary to the seed regions, while the latter are target sites with imperfect seed complementarity (G:U wobbles or mismatches). With the advance of biotechnology, it is accepted that base-pairing can involve both seed regions and non-seed regions (Hafner, Landthaler et al. 2010, Helwak, Kudla et al. 2013, Wang 2014). That is, other types of target sites exist in addition to the canonical and non-canonical target sites. We define unconventional sites as target sites other than the canonical sites. Regardless of the types of target sites, the binding of miRNAs to their target mRNAs during diverse cellular processes may degrade target mRNAs, and/or repress the translation of target mRNAs to proteins (Bartel 2004, Bartel 2009, Wang, Li et al. 2011). Due to such pivotal roles in gene regulation, it is critical to study miRNAs and their target sites.

miRNAs often form modules to regulate target mRNAs (Doench and Sharp 2004, Vella, Choi et al. 2004, Krek, Grun et al. 2005, Saetrom, Heale et al. 2007, Wu, Huang et al. 2010). In this study, a

miRNA module is defined as a group of miRNAs that co-bind a significant number of mRNAs and repress the expression of these common mRNAs significantly more than individual miRNAs in the module (Sections 2.2-2.3). Several studies show that miRNA modules synergistically control the expression of their common target mRNAs (Doench and Sharp 2004, Vella, Choi et al. 2004, Krek, Grun et al. 2005, Saetrom, Heale et al. 2007, Wu, Huang et al. 2010). For instance, miR-124, miR-375, and let-7b form a miRNA module that coordinately regulates the gene Mtpn in a murine pancreatic cell line (Krek, Grun et al. 2005). The distance between adjacent target sites of miRNAs in the same miRNA modules may play critical roles in target mRNA down-regulation (Doench and Sharp 2004, Saetrom, Heale et al. 2007). According to multiple past experiments, the distance of miRNA target sites for optimal down-regulation of target mRNAs is between 13 and 35 nucleotides (Doench and Sharp 2004, Kloosterman, Wienholds et al. 2004, Vella, Choi et al. 2004, Brennecke, Stark et al. 2005). Saetrom et al. computationally showed that miRNA target sites within 130 nucleotides are more likely conserved than more distant sites in 3' UTRs (Saetrom, Heale et al. 2007), suggesting that target sites of miRNAs may need to be within certain ranges to be functional. Several studies also defined miRNA modules by harnessing predicted miRNA target sites in 3' UTRs and the co-expression relationship of target mRNAs of the same miRNAs (Jayaswal, Lutherborrow et al. 2011, Zhang, Li et al. 2011, Bryan, Terrile et al. 2013).

Despite various studies mentioned above, our understanding of miRNA modules is still rudimentary. To our knowledge, all published large-scale studies on miRNA modules so far are based on computationally predicted miRNA target sites in 3' UTRs. However, 3' UTRs only account for a small portion of potential miRNA target site residing regions (Hafner, Landthaler et al. 2010, Helwak, Kudla et al. 2013). Moreover, even the most well-known target site prediction methods

currently produce a significant fraction of false positive target sites (Witkos, Koscianska et al. 2011). In addition, in defining miRNA modules, few computational studies require the higher down-regulation of target gene expression by miRNA modules than that by subsets of miRNAs contained in the modules (Wu, Huang et al. 2010). Therefore, although we have gained basic insight into miRNA modules from previous studies, our understanding of miRNA modules may be biased and limited.

In this study, we employed experimentally determined instead of computationally predicted target sites to study miRNA modules. We predicted 181 miRNA modules and 306 potential miRNA modules. We analyzed binding energy, location, and distances of target sites in these predicted miRNA modules. We observed that target sites of these predicted modules were in general weaker compared with target sites not bound by miRNA modules. We also discovered that miRNAs in predicted modules preferred to bind only unconventional target sites, instead of only canonical sites, or a mixture of canonical and unconventional sites. Contrary to a previous study (Saetrom, Heale et al. 2007), we noticed that most target sites of miRNAs from the same modules were not within the range of 10 to 130 nucleotides. Interestingly, the distance of target sites bound by miRNAs in the same modules was shorter when miRNA modules bound unconventional instead of canonical sites. Our study sheds new light on miRNA binding, which will likely advance our understanding of miRNA regulation.

3.1.2 Material and methods

3.1.2.1 The Clash Data and Gene Expression Data

We downloaded the experimentally-determined miRNA target sites from the crosslinking, ligation, and sequencing of hybrids (CLASH) experiments (Helwak, Kudla et al. 2013). In these experiments, a miRNA and one of its interacting mRNA target sites were ligated and sequenced as a chimeric segment. The sequenced chimerical segments for multiple miRNAs and their target mRNAs were then separated into segments from miRNAs and segments from their target mRNAs, which provided the information of which miRNAs targeted which region of their target mRNAs. We obtained 18514 high-confidence miRNA-mRNA interactions in HEK293 cells. These interactions involved 399 miRNAs, 7390 mRNAs, 4130 canonical sites, 10300 non-canonical sites, and 14384 unconventional sites. Among these target sites, 1034 (5.6%), 11367 (61.4%), and 6096 (32.9%) of them were within 5' UTRs, CDSs, and 3' UTRs, respectively. The remaining 17 target sites were not within any annotated mRNA.

We also downloaded mRNA expression data from (Schmitter, Filkowski et al. 2006). The expression data before and after AGO2 protein knocking down in HEK293 was used to measure the down-regulation of miRNAs and miRNA modules. This was because the knock-down of the AGO2 protein basically prevented the functions of all miRNAs, as AGO2 is an essential component of the RNA-induced silencing complex that recognizes target mRNAs and loads miRNAs to target mRNAs (Bartel 2009).

3.1.2.2 The Pipeline to Predict miRNA Modules

We developed the following pipeline to predict miRNA modules (Figure 3.1): Starting from 18514 experimentally validated target sites, we modified the ChIPModule approach (Ding, Cai et al. 2013)

to discover groups of miRNAs that co-bind at least S mRNAs; Next, we applied the binomial test to assess the statistical significance of every group of miRNAs identified above. Each significant group of miRNAs was called a miRNA module candidate; Finally, we predicted miRNA modules based on hypergeometric testing. A miRNA module was defined as a module candidate that significantly decreased the expression of their common mRNA targets than individual miRNAs contained in this candidate did. The details were in the following two sections.

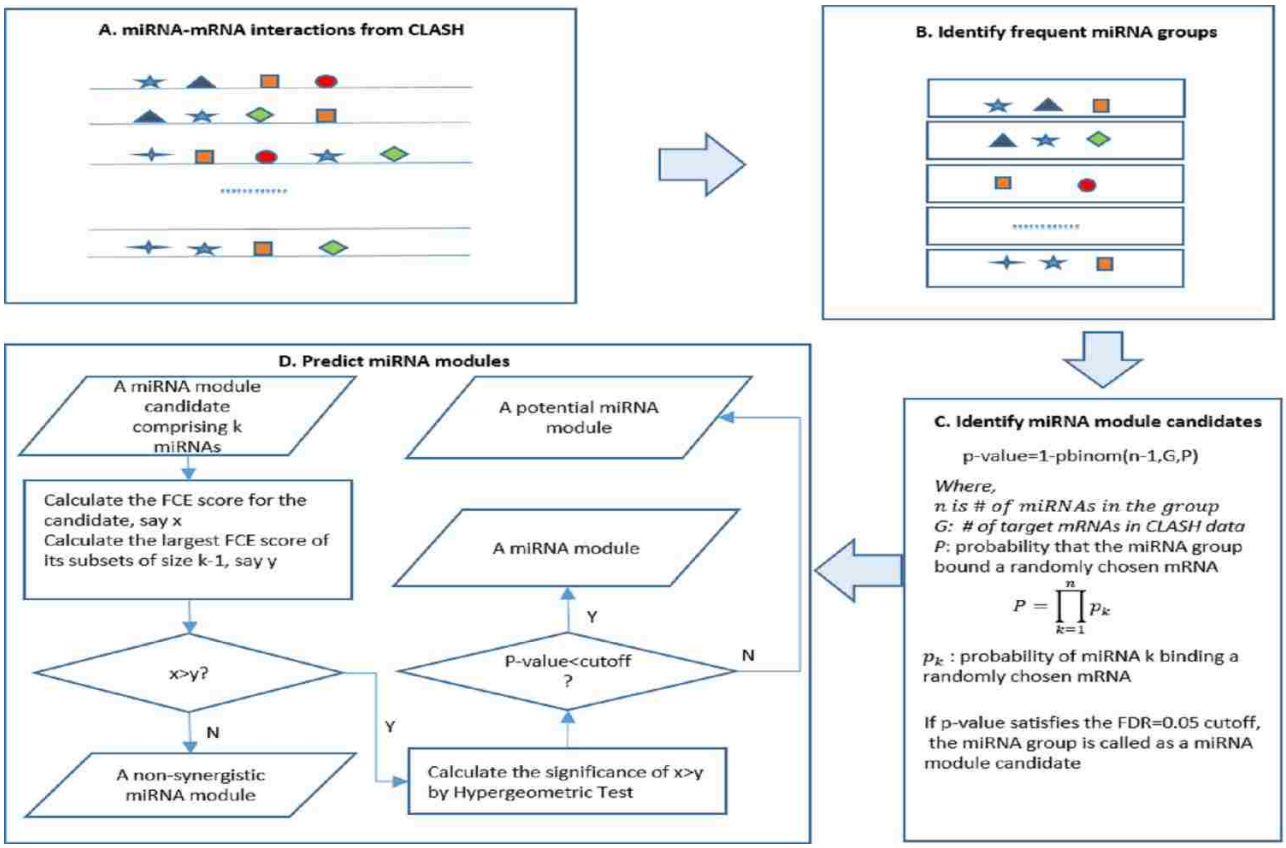

Figure 3-1 The pipeline to predict miRNA modules. **A**. MiRNA-mRNA interaction data from CLASH. Each line represents a target mRNA, each box represents a miRNA target site, with different shapes representing different miRNAs. **B**. Identify miRNA groups with their target sites frequently co-occurring in common mRNAs. **C**. Identify miRNA module candidates by binomial tests. **D**. Predict miRNA modules based on hypergeometric tests.

3.1.2.3 Predicting miRNA Module Candidates

We modified ChIPModule (Ding, Cai et al. 2013) to discover miRNA module candidates. ChIPModule was developed to discover significantly co-occurring binding sites of a group of transcription factors (TFs) in input sequences. In brief, with all known motifs in a database, ChIPModule defines putative TF binding sites (TFBSs) in input sequences. It then identifies TF groups of variable sizes with frequently co-occurring TFBSs in input sequences by an effective tree-based approach. Finally, ChIPModule assesses the statistical significance of each TF group with frequent co-occurring TFBSs by Poisson clumping heuristic (Hu, Hu et al. 2008) and output significant TF groups as TF modules. Because of its superior performance to other methods in TF module discovery (Ding, Cai et al. 2013), we applied a modified version of ChIPModule to miRNA module candidate discovery.

We modified ChIPModule in two aspects. One was that we considered experimentally determined miRNA target sites instead of putative TFBSs (Ding, Cai et al. 2013). The other was that we calculated statistical significance of the co-occurrence of a group of miRNAs differently. Assume we observed a group of k miRNAs, all of which bound n of the 7390 mRNAs. We assessed the statistical significance of this miRNA group as follows: First, for each miRNA, we calculated its probability to have a target site in a randomly chosen mRNA, which was the ratio of the number of mRNAs containing the CLASH target sites of this miRNA to 7390. Second, we calculated the probability that this group of miRNAs bound a randomly chosen mRNA by assuming each miRNA bound the mRNA independently. That is, this probability was measured as the product of the k probabilities that each of the k miRNAs bound the mRNA. Finally, we calculated the binomial tail probability of observing n of the 7390 mRNAs containing target sites of all k miRNAs in this group as the

statistical significance of this group of miRNAs. With the tail probabilities for all groups of miRNAs that frequently co-bind their common mRNA targets, we applied the Q-value software (Storey and Tibshirani 2003) to output significant miRNA groups so that the false discovery rate (FDR) was controlled. The significant miRNA groups were considered as miRNA module candidates (Figure 3-1).

We applied this modified approach to output miRNA module candidates, with S=10 and FDR=0.05. S=10 referred to the requirement that all miRNAs in a module candidate shared at least 10 common target mRNAs. We also tried S=20 to predict candidates. All predicted candidates with S=20 were a subset of the predicted candidates with S=10, which suggested S=20 may be too stringent. We thus reported our analyses with S=10.

3.1.2.4 Identifying miRNA Modules

Given a module candidate comprising k miRNAs, all of which bound the same n of the 7390 mRNAs, we determined whether it was a miRNA module by the following procedure (Figure 3-1). First, we calculated a fold change extent (FCE) score for the candidate and all its subsets with k-1 miRNAs. The FCE score of a group of miRNAs was defined as the fraction of their common target mRNAs with fold changes larger than a pre-defined cutoff D, when the expression levels of genes before and after AGO2 protein knock-down were compared. We used three cutoffs of D, corresponding to the 99%, 95%, and 90% quartile of the distribution of the fold changes of the 7390 mRNAs. Second, we checked whether the FCE score of this miRNA module candidate was larger than that of any subset of size k-1. Third, if this candidate had larger FCE scores than any subset, we assessed its significance of higher down-regulation of target genes than the k subsets by the following hypergeometric testing. Without loss of generality, we assumed that (i) the subset with the

largest FCE score had N common target mRNAs, among which M mRNAs had fold changes larger than D; and (ii) the fold changes of m out of the n target mRNAs of this candidate were larger than D. Under these assumptions, the significance of observing higher down-regulation of targets of the module than that of any subset was measured by a hypergeometric testing tail probability of observing at least m of n targets with large fold changes randomly chosen from the population of N targets with M targets of large fold changes. Finally, we assigned the module type of this candidate. If the significance at step three satisfied the required FDR of 0.05 (Storey and Tibshirani 2003), this candidate was claimed as a synergistic miRNA module. Otherwise, if the FCE score of the candidate was larger than that of any subset at step two, this candidate was a potential synergistic miRNA module. In all remaining cases, this candidate was considered as a non-synergistic miRNA module. The synergistic, potential, and non-synergistic miRNA modules were abbreviated as miRNA modules, potential modules, or non-synergistic modules below, respectively.

With the above procedure, we predicted 193, 181, and 190 miRNA modules, using the 99%, 95%, and 90% quartile of the distribution of fold changes as D, respectively. More than 80% of the predicted miRNA modules with the three choices of D were shared by the three sets. For convenience, we reported our results based on the second choice of D (95%).

3.1.2.5 Validating Predicted Candidates Especially the Predicted miRNA Modules

To assess whether the predicted candidates especially the miRNA modules were likely functional, we studied the overlap of their target genes with known pathway genes or gene sets of known functions. For known pathway genes, we used pathways at http://www.broadinstitute.org/gsea/index.jsp. To evaluate the overlap significance, we used hypergeometric testing (Boyle, Weng et al. 2004). We also checked the order of target sites of different miRNAs in a predicted candidate as in previous

studies (Cai, Hou et al. 2010, Ding, Hu et al. 2014). We also searched literature to see whether a predicted module was supported as well. For each candidate, we searched in Google scholar to retrieve top 20 hits. For every hit, we manually checked whether all miRNAs contained in this candidate were reported to (i) bind the same targets; (ii) be active under the same experimental conditions (e.g. highly co-up-regulated in specific cancers); or (iii) be found in co-transfection experiments. If at least one type of evidence was found, this candidate was considered to be supported by literature.

3.1.2.6 Comparing the Strength of Target Sites

We compared the strength of target sites bound by predicted modules with that by individual miRNAs. The strength was measured by target site binding energy downloaded from the CLASH paper (Helwak, Kudla et al. 2013). The binding energy approximated the interaction strength of a miRNA and one of its target sites. The higher the energy was, the weaker the target site was. In brief, we treated the binding energy of target sites bound by predicted miRNA modules as samples of a random variable X1. We also obtained the samples of another random variable Z, which was the binding energy of the sites in mRNAs that were not bound by any predicted candidate. We then applied Wilcoxon rank-sum test to test the null hypothesis Z>X1 with the alternative hypothesis as Z<=X1 (Wilcoxon 1945). Similarly, we compared the strength of target sites of the potential modules with that of individual miRNAs. We also compared the strength of target sites of individual miRNAs in individual modules with the strength of target sites of the same miRNAs but not bound by any candidate.

3.1.2.7 Analyzing the Preferred Target Site Combinations of miRNA Modules

We investigated the combinations of target sites a (potential) miRNA module preferred to bind. We had three possible combinations: all canonical sites (type 1), a mixture of canonical and unconventional sites (type 2), and all unconventional sites (type 3). For a (potential) miRNA module comprising k different miRNAs, the probability of the type 1 combination of target sites was pk, where p is the fraction of the number of canonical sites in the 18514 target sites. Similarly, the probability of the type 3 combination was (1-p)k. The probability of the type 2 combination was 1-pk - (1-p)k. We assume that we observed this (potential) module targeting n mRNAs, m1, m2, and m3 of which had the type 1, type 2, and type 3 target site combinations, respectively. We calculated the significance of this (potential) module preferring the type 1 combination as the binomial tail probability of observing at least m1 successful experiments in n experiments, each of which had a success rate of pk. Similarly, we calculated the significance of this (potential) module preferring other types of combinations.

We analyzed five different location combinations of target sites: all sites in CDSs (type 1), all sites in 3' UTRs (type 2), all sites in 5' UTRs (type 3), at least one 3' UTR site and another site not from 3' UTR (type 4), and all other sites (type 5). We did similar tests to determine whether a (potential) miRNA module preferred a specific combination of site locations.

3.1.2.8 Inferring Preferred Distance Ranges of Adjacent Target Sites of miRNA Modules

We defined preferred distance ranges of adjacent target sites of a miRNA module as follows: First, we divided the distances of adjacent target sites of a miRNA module into 10 nucleotides long bins. Second, we calculated the p-value of the enrichment of distances in each bin, by assuming the distances were evenly distributed across bins. If the p-value is small (FDR=0.05), the bin was considered as significantly enriched. Third, we extended each significant bin to obtain a region with

the smallest p-value of enrichment and defined this region as a preferred region. More precisely, for a significant bin, say A, we considered combining A with its left neighboring bin. We calculated the p-value of the enrichment of distances in these two bins under the same uniform distribution assumption. Similarly, we considered the two bins comprising A and its right neighboring bin. We then chose the extension with the smaller p-value, for instance, the extension to the left. If this smaller p-value is small (FDR=0.05) and smaller than the p-value of A, we extended A into a preferred region comprising two bins. We repeated this procedure until no more extension could be made. Finally, we reported all non-overlapping preferred regions as the preferred distance ranges of adjacent target sites of this miRNA module. Similarly, we defined the preferred distance ranges of adjacent target sites of other types of candidates.

3.1.3 Results

3.1.3.1 Predicted miRNA Modules Were Supported by Functional Evidence

With FDR=0.05 (Storey and Tibshirani 2003), we discovered 507 miRNA module candidates. Each candidate consisted of 2 to 5 miRNAs, with an average of 2.72 miRNAs. The number of candidates with 2, 3, 4, and 5 miRNAs was 174, 300, 32, and 1, respectively. All miRNAs in a candidate shared at least 10 common mRNA targets.

We investigated whether the 507 candidates significantly down-regulated their target mRNAs more than any subset of the contained miRNAs (Material and Methods). We found 181 candidates down-regulated their target mRNAs significantly more than their contained subsets (FDR=0.05). We considered these 181 candidates miRNA modules. We also noticed that 306 candidates down-regulated their target mRNAs more than their contained subsets, with or without satisfying the

required FDR of 0.05. We considered these 306 candidates potential miRNA modules, which included the above 181 modules. The remaining 201 candidates, which did not down-regulate their target genes more than their subsets, were defined as non-synergistic modules.

We explored whether the 507 candidates, especially the 181 miRNA modules, were functional. We studied the overlap of target mRNAs of a candidate with genes in a known pathway or annotated with a common gene ontology (GO) term, as in previous studies (Ambros 2004, Xu, Li et al. 2011). The rationale was that if a candidate was functional, its target genes likely significantly overlap with genes in a known pathway, or genes annotated with a specific GO term (Ambros 2004, Xu, Li et al. 2011). We found that the function of the majority of the predicted candidates, especially the predicted miRNA modules, was supported (Table 3.1). For instance, the target mRNAs of 68.4% of the 507 candidates significantly shared at least one GO term. To assess the statistical significance of the pathway and GO support, we generated 507 random miRNA groups, each of which consisted of miRNAs randomly chosen from the 399 miRNAs mentioned in the CLASH paper and contained the same number of miRNAs as the corresponding predicted candidate. We found that target mRNAs of only 18 random miRNA groups significantly overlap with genes in a pathway, and target mRNAs of only 6 random miRNA groups shared a GO term (Table 3.1).

In addition, we studied the order of target sites of miRNAs in the same predicted candidates. We found that 34.3% of the 181 miRNA modules, 33.7% of the 306 potential modules, and 29% of the 507 candidates contained miRNA pairs with statistically significant orders (Table 3.1, FDR=0.05). For instance, three miRNAs MIR-222, LET-7B, and MIR-615-3P formed a miRNA module. MIR-615-3P preferred to bind at the 5' of the target sites of MIR-222 (FDR=0.0286), which preferred to

bind at the 5' of the target sites of LET-7B (FDR=6.1E-4). In contrast, no random miRNA group had preferred orders.

We also did a literature search to check whether the predicted candidates was supported. We found that 99 of the 507 candidates and 32 of the 181 miRNA modules were supported by literature (Table 3.1). By comparison, we did literature search for the 507 random miRNA groups and found that 11 groups were supported.

Table 3.1 Support of the predicted miRNA module candidate

Module types	# (%) of combinations supported by Pathway	# (%) of combinations supported by GO	# (%) of combinations supported by literature	Significance of literature support	# (%) of combinations supported by order	Total # (%) of combinations supported
181 synergistic modules	125 (69.0%)	165 (91.2%)	32 (17.7%)	1.87e-12	62 (34.3%)	178 (98.3%)
306 possible modules	211 (69.0%)	274 (89.5%)	57 (16.8%)	0	103 (33.7%)	298 (97.4%)
201 non-synergistic modules	136 (67.7%)	174 (86.6%)	42 (20.9%)	0	44 (21.9%)	194 (96.5%)
507 predicted candidates	347 (68.4%)	448 (88.4%)	99 (19.5%)	0	147 (29.0%)	492 (97%)

The supporting rate of 11/100 was likely over-estimated, as the 399 miRNAs were known to be active under the same conditions and may actually play certain roles together. Even given the over-

estimated supporting rate of random miRNA combinations, the chance of observing the number of supported candidates was 0 (Table 3.1).

3.1.3.2 Target Sites of miRNA Modules Were Weaker Compared with Other Target Sites

We compared the strength of target sites bound by predicted (potential) miRNA modules with that by individual miRNAs. Target sites of the predicted miRNA modules had significantly higher energy than target sites bound by individual miRNAs (Wilcoxon rank-sum p-value 3.0E-19). Similarly, this was true for target sites of the potential miRNA modules (p-value 4.6E-15). Target sites of (potential) miRNA modules were thus weaker than target sites not bound by any of the 507candidates.

We also compared the strength of target sites of the predicted modules and the potential modules with that of non-synergistic modules. Target sites of the 181 miRNA modules had significantly higher energy than target sites bound by the 201 non-synergistic modules (p-value 1.6E-56). Similarly, target sites of the 306

potential modules had significantly higher energy than those bound by the 201 non-synergistic modules (p-value 1.7E-87).Therefore, target sites of the (potential) miRNA modules were weaker than those of the 201 non-synergistic modules, which implied that there was a difference between candidates that down-regulated target mRNAs more than their subsets and candidates that did not.

We further compared the strength of target sites of individual miRNAs bound by (potential) miRNA modules with that of target sites of the same miRNAs not bound by any of the 507 candidates. For 42 of the 56 miRNAs in the 181 miRNA modules, their target sites were significantly weaker compared with their target sites not bound by any of the 507 candidates (FDR=0.05). On the contrary, for only

4 of the 56 miRNAs, their target sites were significantly stronger than those not bound by any candidate (FDR=0.05), suggesting that for the majority of miRNAs, target sites bound by modules were not as strong as target sites not bound by any candidate. A similar observation was made for the 306 potential modules, for which target sites of 51 of the 68 miRNAs bound by modules were weaker and target sites of only 7 of the 68 miRNAs bound by modules were stronger (FDR=0.05). Therefore, consistent with the above pooled analyses of target sites of all (potential) modules, target sites of the majority of miRNAs bound by individual (potential) modules were weaker (Figure 3-2).

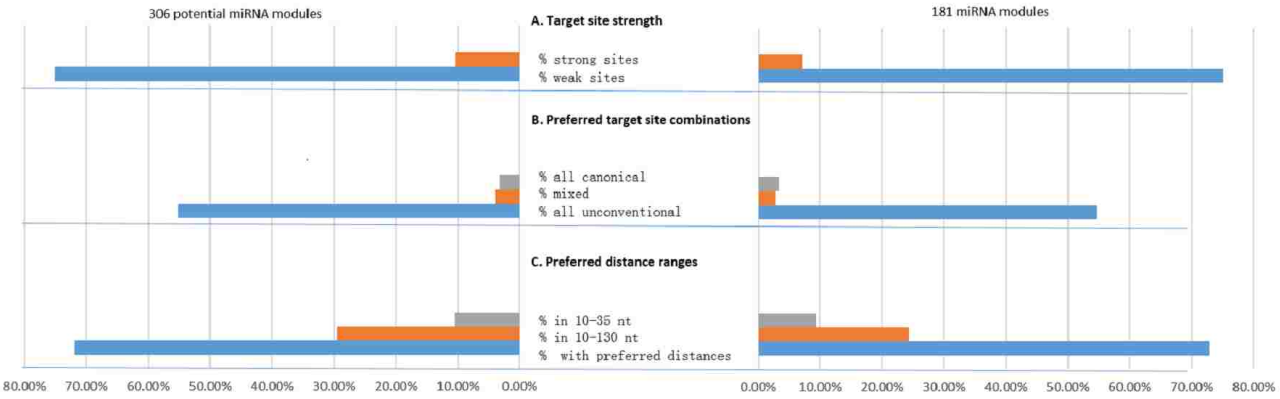

Figure 3-2 The target site strength, site combinations, and distances of adjacent sites preferred by miRNA modules. **A**. Percentages of miRNAs in modules preferring sites of different strengths. **B**. Percentages of preferred target site combinations. **C**. Percentages of preferred distance ranges.

3.1.3.3 Most miRNA Modules Preferred to Bind Unconventional Target Sites

We investigated the preferred combinations of canonical and unconventional target sites of a (potential) miRNA module. For the 181 miRNA modules, 99, 5, and 6 miRNA modules preferred target sites composed of all unconventional sites (type 1), a mixture of canonical and unconventional sites (type 2), and all canonical sites (type 3), respectively (FDR=0.05). Similarly, for the 306 potential miRNA modules, 169, 12, and 10 potential miRNA modules preferred target sites of types 1, 2, 3, respectively (FDR=0.05). It was thus evident that more than half of the (potential) miRNA

modules preferred to bind unconventional sites instead of a mixture of canonical and unconventional sites.

The above analyses only showed that most (potential) miRNA modules preferred to bind mRNAs containing only unconventional sites. It was not clear whether target sites in the majority of mRNA targets of a (potential) miRNA module were composed of only unconventional sites. We thus checked the (potential) miRNA modules which had more than 70% of mRNA targets that contained only unconventional sites. We found that 116 of the 181 (64.1%) miRNA modules had more than 70% of their mRNA targets with only unconventional sites, and 198 of the 306 (64.7%) potential miRNA modules had more than 70% of mRNA targets with only unconventional sites. Therefore, the combination of all unconventional target sites was the most dominant combination found in target mRNAs of most (potential) miRNA modules.

Since unconventional sites included non-canonical sites and other sites, we explored whether miRNA modules prefer to bind non-canonical sites. For the 181 significant modules, only 52 (28.7%) modules significantly preferred all non-canonical sites. Moreover, only 4 modules had more than 70% of mRNA target with only non-canonical sites. Similarly, for 306 potential synergistic modules, only 82 (26.8%) modules significantly preferred all non-canonical sites. Only 5 modules had more than 70% of mRNA target with only non-canonical sites. It was thus evident that miRNA modules preferred unconventional sites instead of non-canonical unconventional sites (Figure 3-2).

3.1.3.4 Most miRNA Modules Preferred to Bind the First or the Last Exons

We studied the location combinations of target sites of (potential) miRNA modules (Material and Methods). We found that most (potential) miRNA modules preferred to bind target sites in CDSs

instead of a mixture of CDSs and UTRs. For the 181 miRNA modules, 127, 7, 2, and 2 miRNA modules preferred target sites all in CDSs (type 1), all in 3' UTRs (type 2), all in 5' UTRs (type 3), and in both 3' UTRs and other locations (type 4), respectively (FDR=0.05). Similarly, for the 306 potential miRNA modules, 210, 16, 3, and 2 potential miRNA modules preferred target sites of types 1, 2, 3, and 4, respectively. In contrast, 1, 0, 0, and 0 of the 507 random miRNA groups preferred target sites of types 1, 2, 3, and 4, respectively.

Because the majority of (potential) miRNA modules preferred to bind only in CDSs, we further investigated whether target sites in an mRNA target of a (potential) module were always in the same exons, adjacent exons, or others. We found that the majority of the (potential) miRNA modules prefer to bind the same exons. For instance, for 150 of the 181 miRNA modules, more than 50% of their target sites in a target mRNA were in the same exon. Similarly, for 256 of the 306 potential modules, more than 50% of their target sites in a target mRNA were in the same exon. In contrast, the 507 random miRNA groups did not have exon binding preference, of which only 14 random miRNA groups had more than 50% of their target sites in the same exon in an mRNA.

Since most target sites in a target mRNA of a (potential) module were in the same exon, we also studied whether the target sites of a (potential) module preferred a specific type of exons. Indeed, we found that the first and the last exon of the target mRNAs were preferred to be bound by the (potential) miRNA modules. For instance, 64.9% of target sites of the 181 miRNA modules that were in the same exons were in either the first exons or the last exons. Similarly, 67.3% of target sites of the 306 potential miRNA modules that were in the same exons were in either the first exons or the

last exons. Therefore, we concluded that most (potential) miRNA modules preferred to bind one exon in the target mRNAs, either the first or the last exons.

3.1.3.5 MiRNA Modules Preferred Target Sites within Certain Ranges

With the experimentally determined target sites, we studied the preferred distance range of the adjacent target sites of miRNA modules. We defined the preferred distance ranges of target sites of the 181 miRNA modules, 306 potential miRNA modules, 202 non-synergistic modules, and the 507 module candidates, respectively (Table 3.2). For each of the four types of miRNA combinations, more than 70% of combinations had preferred distance ranges. The distribution of the preferred distance ranges of the 181 miRNA modules and that of the 306 potential modules were more similar to each other, compared with that of the 202 non-synergistic modules (Figure 3-3). For all four types of miRNA combinations with preferred distance ranges, more than 90% of combinations in each type had all preferred distance ranges <360 nucleotides (Figure 3-3 and Table 3.2). In contrast, 20 (3.9%) of the 507 random miRNA groups had preferred distance ranges, 10 (50.0%) of which had preferred distance ranges <360 nucleotides.

We tested whether two previously known preferred ranges, 10-130 nucleotides and 13-35 nucleotides, were enriched. For 44 (24.3%) miRNA modules, 90 (29.4%) potential miRNA modules, and 69 (34.3%) non-synergistic modules, the 10-130 range was enriched (Binomial test, FDR=0.05). For 17 (9.4%) miRNA modules, 32 (10.5%) potential miRNA modules, and 16 (8.0%) non-synergistic modules, the 13-35 range was enriched (Binomial test, FDR=0.05). If we pooled the distances of target sites of all 181 miRNA modules together, the 10-130 and 13-35 ranges were enriched (p-value=0 in both cases). Similarly, the two ranges were enriched in distances of target sites of the potential modules, the non-synergistic modules, and the 507 module candidates. Although

the two ranges of distances were enriched, the majority of distances of adjacent target sites of modules or module candidates were not within the two ranges (Table 3.2). For instance, in the 181 miRNA modules, more than 74.1% of distances of adjacent target sites of miRNA modules were longer than 130 nucleotides. Moreover, 85 (47.0%) miRNA modules had preferred distance ranges larger than 130 nucleotides (Table 3.2 and Figure 3-3). Future work may need to investigate how miRNAs with target sites in such large distance ranges interact.

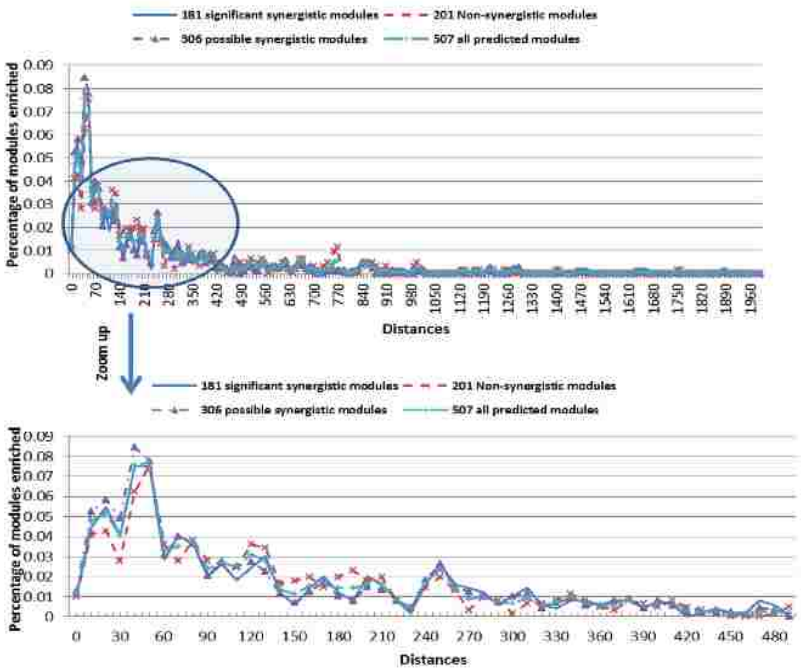

Figure 3-3 MiRNA modules preferred target sites within certain ranges

3.1.3.6 Adjacent Unconventional Target Sites of miRNA Modules Preferred Shorter Distances than Other Types of Adjacent Target Sites

We studied the difference between the distances of adjacent unconventional target sites of a module and those of other types of adjacent target sites of the same module. For every predicted miRNA module, we collected the distances of adjacent unconventional target sites in each target mRNA. That is, we only considered target mRNAs that contained only unconventional sites of this module. We also collected distances of other target sites of this module. For 150 of the 181 miRNA modules, we

had at least 5 distances collected for each of the two types of distances. We tested the null hypothesis that the distances of adjacent unconventional sites were shorter than those of other types. We only rejected the null hypothesis in 5 cases based on Wilcoxon rank-sum test (FDR=0.05). We further tested the null hypothesis that the distances of adjacent unconventional sites were longer than those of other types. We rejected the null hypothesis in 31 cases (FDR=0.05). We concluded that in 6 times more cases, the distances of adjacent unconventional target sites seemed shorter than that of other types of adjacent target sites.

The sample size may be too small for the above analyses using individual miRNA modules. Therefore, we considered distances of adjacent unconventional sites of all miRNA modules together instead of individual miRNA modules. For the null hypothesis that the distances of adjacent unconventional sites were shorter than those of other types, we accepted the null hypothesis (p-value>0.99). For the null hypothesis that the distances of adjacent unconventional sites were longer than those of other types, we rejected the null hypothesis (p-value=9.7E-14). This pooled analysis showed that the distances of adjacent unconventional sites were indeed shorter.

We also did similar analyses for the 306 potential modules. We obtained similar results. That is, for the analysis based on individual potential modules, in 3 times more cases (41 versus 12), the distances of adjacent unconventional sites were shorter than the distances of other types of sites. In the pooled analysis, for the null hypothesis that the distances of adjacent unconventional sites were shorter than those of other types, we accepted the null hypothesis (p-value>0.99). For the null hypothesis that the distances of adjacent unconventional sites were longer than those of other types,

we rejected the null hypothesis (p-value=5.0E-16). Therefore, it was evident that adjacent unconventional target sites preferred shorter distances than other types of adjacent target sites.

Table 3.2 Preferred distance ranges of adjacent target sites of miRNA combinations

Module types	% of modules with preferred distance ranges	% of modules with preferred distance ranges overlapping with 10-130nt	% of modules with preferred distance ranges overlapping with 13-35nt	% of adjacent distances <130nt	% of modules with preferred distance ranges>130nt	% of modules with preferred distance ranges< 360nt
181 significant synergistic modules	132/181 =72.9%	44/181 =24.3%	17/181 =9.4%	25.9%	85/181 =47.0%	119/132 =90.2%
306 possible synergistic modules	220/306 =71.9%	90/306 =29.4%	32/306 =10.5%	28.1%	142/306 =46.4%	198/220 =90.0%
201 non-synergistic modules	148/201 =73.6%	69/201 =34.3%	16/201 =8.0%	27.4%	105/201 = 52.2%	136/148 =91.9%
507 all predicted modules	368/507 =72.6%	159/507 =31.4%	48/507 =9.5%	27.8%	247/507 =48.7%	334/368 =90.1%
507 random miRNA groups	20/507 =3.9%	4/507 =0.8%	2/507 =0.4%	10.7%	16/507 =3.2%	10/20 =50.0%

3.1.4 Discussion

We studied miRNA modules based on experimentally determined miRNA target sites. We predicted 181 miRNA modules and 306 potential miRNA modules. We demonstrated that miRNA modules preferred to bind weak sites and favored a combination of all unconventional sites. We also observed that miRNA modules preferred to bind in CDSs and favored the first and the last exons. We confirmed that more than 70% of miRNA modules bound sites within specific ranges, with

enrichment in two previously known ranges. However, many more adjacent sites bound by miRNA modules were >130 nucleotides apart. We further showed that unconventional target sites of miRNA modules were often within shorter distances than other combinations of target sites. Our study shed new light on miRNA binding.

The majority of adjacent target sites of miRNA modules were >130 nucleotides apart, which contradicted with previous observations (Doench and Sharp 2004, Kloosterman, Wienholds et al. 2004, Vella, Choi et al. 2004, Brennecke, Stark et al. 2005, Saetrom, Heale et al. 2007). To understand what resulted in different observations, we focused on target sites of the 181 miRNA modules in 3' UTRs. We found even when we considered only target sites in 3' UTRs, more than 75% of adjacent target sites of miRNA modules were >130 nucleotides apart. We also predicted miRNA module candidates using only the 6096 CLASH target sites in 3' UTRs and then studied the distances of adjacent target sites of these candidates. We still observed that the majority of adjacent target sites of these candidates were >130 nucleotides apart. Therefore, the different observations were unlikely because we used target sites in entire mRNA regions while previous studies used only target sites in 3' UTRs. Instead, it may be due to the small number of experimentally determined sites in previous experimental studies and the limited quality of predicted sites in the previous computational study, compared with the 18514 high-quality experimentally determined sites we used.

We predicted (potential) miRNA modules on the condition that they down-regulated target genes significantly more than some of their miRNA subsets. We further checked whether these (potential) modules down-regulated their target genes significantly more than any subset contained in the

modules. We confirmed that for all (potential) miRNA modules, their target genes were significantly more down-regulated than the target genes of any of their subsets.

We discovered 201 non-synergistic modules. The non-synergistic modules may also play important roles in regulating target genes, as supported by GO and pathway analyses, order preference, and the literature. Moreover, these non-synergistic modules may be competitive miRNA modules that are worth further investigation (Khan, Betel et al. 2009).

3.2 TarPmiR: a new approach for microRNA target site prediction

3.2.1 Background

The prediction of microRNA (miRNA) target sites is critical in understanding miRNA function and their involvement in various biological processes (Lewis, Shih et al. 2003). MiRNAs are short noncoding RNAs that bind and regulate their target mRNAs in a variety of biological processes, such as cell development, differentiation, proliferation, and apoptosis pathways (Sassen, Miska et al. 2008, Schanen and Li 2011). The binding of miRNAs to their target mRNAs degrades the target mRNAs and/or prevents the target mRNAs from being translated into proteins, and thus modulates gene expression at the post-transcriptional level (Friedman, Farh et al. 2009, Muljo, Kanellopoulou et al. 2010, Axtell, Westholm et al. 2011, Wang, Li et al. 2011). By identifying miRNA target sites, the target mRNAs and the potential functional roles of miRNAs may thus be discovered.

Several features are commonly believed to be important for predicting miRNA target sites. Among them, seed match, the exact sequence matching between the positions 2 to 7 of an miRNA and a segment of 6 nucleotides (nt) long in target mRNAs, has been reported to be essential for miRNA-mRNA binding (Brennecke, Stark et al. 2005). Accessibility, which measures how likely a region in an mRNA sequence is "open" or accessible for an miRNA to bind, is well known to be important for functional miRNA-mRNA binding (Kertesz, Iovino et al. 2007). In addition, other features such as AU content (Grimson, Farh et al. 2007), folding energy (Enright, John et al. 2004, Grimson, Farh et al. 2007, Yousef, Jung et al. 2007), and conservation (Helwak, Kudla et al. 2013) are also regarded as informative indicators of functional miRNA-mRNA bindings.

Dozens of tools for miRNA target site prediction have been developed in the past decade, based on different subsets of the aforementioned features (Peterson, Thompson et al. 2014). For instance, miRanda (Enright, John et al. 2004) utilizes the features of seed match, conservation and free energy for target site prediction. TargetScan (Grimson, Farh et al. 2007, Friedman, Farh et al. 2009) uses seed match, pairing of mRNAs with 3' of miRNAs, local AU content, etc, for target site identification. In addition to these traditional miRNA target site prediction tools, recently, several tools based on next-generation sequencing technologies have been developed (Vejnar and Zdobnov 2012, Chou, Lin et al. 2013, Wang, Xie et al. 2014). For instance, miRTarCLIP (Chou, Lin et al. 2013) identifies miRNA target sites from the data generated by high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) experiments (Licatalosi, Mele et al. 2008, Chi, Zang et al. 2009) and photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) experiments (Hafner, Landthaler et al. 2010).

Despite the existence of dozens of computational methods, computational identification of miRNA target sites remains a challenging problem partially due to our limited understanding of the characteristics of miRNA target sites. For instance, although matching seed is not always sufficient for a functional miRNA-mRNA interaction (Brennecke, Stark et al. 2005, Didiano and Hobert 2006), it has been thought to be necessary for most animal miRNA-mRNA binding. However, studies have shown non-canonical pairings that allow G:U wobbles and even mismatches can be functional (Brennecke, Stark et al. 2005, Didiano and Hobert 2006). Recent cross-linking ligation and sequencing of hybrids (CLASH) experiments (Helwak, Kudla et al. 2013) have further shown that seed match, including canonical and non-canonical seed-matching, is not required for certain miRNA-mRNA interactions.

The CLASH experiments (Helwak, Kudla et al. 2013) provide an unprecedented opportunity to advance our understanding of miRNA target sites and to develop better computational methods for miRNA target site prediction. Compared with other high-throughput experimental approaches such as HITS-CLIP (Licatalosi, Mele et al. 2008, Chi, Zang et al. 2009) and PAR-CLIP (Hafner, Landthaler et al. 2010) that identify miRNA target sequences only, CLASH experiments provide both miRNAs and their corresponding target sequences. With thousands of target sequences for dozens of miRNAs in one CLASH experiment, new features of miRNA target sites may be inferred and better computational methods for miRNA target site prediction may be developed.

In this study, we developed a new approach for miRNA target site prediction called TarPmiR (Target Prediction for miRNAs). TarPmiR applies a random-forest-based approach to integrate six conventional features and seven new features to predict miRNA target sites. These features were learned from the only CLASH dataset in mammal that is made publically available by Helwak, et al. (Helwak, Kudla et al. 2013). By cross-validation, we showed that TarPmiR had an average recall of 0.543 and an average precision of 0.181. Tested on three independent datasets, including two human PAR-CLIP datasets and one mouse HITS-CLIP dataset, we demonstrated that TarPmiR identified more than 74.2 % of known miRNA target sites in each dataset. Compared with three existing approaches, we found that TarPmiR is superior to existing approaches, in terms of both higher recall and higher precision. The TarPmiR method is implemented in a python package, which is freely available at http://hulab.ucf.edu/research/projects/miRNA/TarPmiR/.

3.2.2 Materials and Methods

3.2.2.1 Training and testing data

We downloaded 18514 miRNA target sites of 399 miRNAs from CLASH experiments (Helwak, Kudla et al. 2013). These target sites were considered as positive target sites. We also generated 18514 corresponding negative or "false" target sites in a manner similar to a previous study (Li, Kim et al. 2014), with the following criteria: (1) A positive site and its corresponding negative site are on the same mRNA; (2) The positive and its corresponding negative site has similar CG dinucleotide frequency; (3) The positive and its corresponding negative site has similar number of the nucleotide G; (4) A negative site does not overlap with any positive site; (5) With multiple candidate negative sites in an mRNA, select the one with the lowest folding energy.

We performed cross-validation to determine which machine learning method to be used in TarPmiR and to assess the accuracy of TarPmiR. To determine which method to be used, we randomly chose 10000 positive sites and 10000 negative sites for training and the remaining positive and negative sites for testing. We repeated this process five times and selected the method with the F2 scores. To test TarPmiR, we used the same five training datasets. For a corresponding testing dataset, we input the mRNA sequences that contain the corresponding remaining 8514 positive sites and the remaining 8514 negative sites for testing. The final model used to predict miRNA target sites by TarPmiR in this study was trained using the first set of randomly chosen 10000 positive sites and 10000 negative sites.

We also collected two independent PAR-CLIP datasets from the human HEK293 cell line for testing. PAR-CLIP datasets were used because a large number of potential miRNA target regions called

crosslink-centered regions (CCRs) could be obtained from PAR-CLIP. CCRs were considered as positive target sites. One PAR-CLIP dataset with 17310 CCRs was from (Hafner, Landthaler et al. 2010). Only 16041 of these CCRs were able to be mapped to mRNAs and resulted in 10023 target mRNAs. In this dataset, 60 miRNAs accounted for more than 90% of total miRNA reads and 120 miRNAs accounted for 99% of total miRNA reads. In other words, depending on the cutoff to define active miRNAs, there were mainly 60 or 120 miRNAs related to these 17310 CCRs. The other PAR-CLIP dataset with 44497 CCRs was obtained from (Kishore, Jaskiewicz et al. 2011). Only 43251 of the 44497 CCRs were able to be mapped to mRNAs and resulted in 17794 target mRNAs. Same as the first PAR-CLIP dataset, depending on the cutoff to define active miRNAs, there were mainly 60 or 120 related miRNAs in this dataset.

To test TarPmiR on general datasets, we also compared the TarPmiR predictions with the experimentally validated miRNA targets by general methods in TarBase 6.0 (Vergoulis, Vlachos et al. 2012) There are 15616 POSITIVE TarBase 6.0 miRNA-mRNA interactions in human. We choose the top 100 and 50 miRNAs, which have the largest number of interactions in TabBase 6.0 for further analyses. The rationale to choose top miRNAs is that we have more experimentally validated target mRNAs of these miRNAs and thus can assess the accuracy of TarPmiR and other tools better. The top 100 and 50 miRNAs in TarBase 6.0 accounts for 15552 (99.7%) and 15386 (98.5%) of human TarBase 6.0 interactions, respectively. There are 9869 and 9823 mRNAs associated with these 100 and 50 top miRNAs, respectively. We ran TarPmiR and other tools with the 100 or 50 miRNAs and the corresponding mRNAs they interact as input to predict miRNA target sites.

In addition to the human datasets, we collected an independent HITS-CLIP dataset from the mouse cortex cell (Chi, Zang et al. 2009). This dataset provided an Argo-miRNA-mRNA ternary interaction map related to 20 miRNA families, 2953 mRNAs and 11080 miRNA-mRNA interactions. We further downloaded the corresponding 119 miRNAs from the 20 miRNA families from miRBase (Griffiths-Jones, Grocock et al. 2006).

3.2.2.2 Potential features considered

We considered the following 18 features of miRNA target sites in miRNA-mRNA duplexes: (1) Folding energy; (2) Seed match; (3) Accessibility; (4) AU Content; (5) Stem conservation; (6) Flanking conservation; (7) Difference between stem and flanking conservation; (8) m/e motif; (9) The total number of paired positions; (10) The length of the target mRNA region; (11) The length of the largest consecutive pairs; (12) The position of the largest consecutive pairs relative to the miRNA 5'; (13) The length of the largest consecutive pairs allowing 2 mismatches; (14) The position of the largest consecutive pairs allowing 2 mismatches; (15) The number of paired positions at the miRNA 3' end, where 3' miRNA end was defined as the last 7 positions of the miRNA; (16) The total number of paired positions in the seed region and the miRNA 3' end; (17)The difference between the number of paired positions in the seed region and that in the miRNA 3' end; and (18) Exon preference (Ding, Li et al. 2015). The first seven features had been used in existing tools (Peterson, Thompson et al. 2014), we thus considered them as conventional features. Remaining features that had not been commonly used by miRNA-target prediction tools were defined as 'new' features.

The detailed definition of all 18 features and how to calculate their values are provided in the http://hulab.ucf.edu/research/projects/miRNA/TarPmiR/download/Supp/S1.docx. We briefly explain the m/e motif feature here, as it is not as self-evident as others. The m/e motif describes how different

positions in miRNAs match the corresponding positions in target sites. Here two positions match means that nucleotides at the two positions are complement to each other. For instance, nucleotides at positions in miRNA seed regions tend to match the nucleotides at the corresponding positions in target sites and nucleotides at positions in other miRNA regions tend to form mismatches or bulges with the corresponding positions in target sites. We thus have a sequential pattern composed of two letters "m" and "e" to describe preferred matching and non-matching positions, respectively. To calculate the m/e scores, for each position in miRNAs, we calculate a probability p_i that this position matches the corresponding position in target sites by using all positive target sites in the training dataset. The m/e motif score of a potential target site is calculated as $score = \frac{1}{x}\sum_{i=1}^{x} \log p_i$, where x is the length of the miRNA and x is smaller than 24.

3.2.2.3 Four computational methods for feature selection

Not all of the aforementioned 18 features are effective for target site prediction. To select important features, we applied the following four machine learning methods: step-wise logistic regression (Ralston and Wilf 1960), least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996), randomized logistic regression (Meinshausen and Bühlmann 2010) and random forests (Svetnik, Liaw et al. 2003). The step-wise logistic regression repeatedly eliminates the least significant feature until all significant features are found, which is performed by using the GLM package in R (http://data.princeton.edu/R/glms.html). LASSO constructs a linear model and shrinks the coefficients of non-important features to zero. All features with non-zero regression coefficients are 'selected' as important features. We used the glmnet package (http://cran.r-project.org/web/packages/glmnet/index.html) in R for the LASSO analysis. The randomized logistic regression randomly chooses a portion of the training samples and performs the logistic regression to

select significant features. It repeats this procedure many times and counts the number of times each feature is selected, which is regarded as the importance of the features. The randomized logistic regression was performed with the scikit-learn package (http://scikit-learn.org/stable/) in python. The random forests method grows many classification trees and assigns a new object to the class most trees vote for this object. We used the random forest model from sklearn package (http://scikit-learn.org/stable/) in python. Each of the four methods has been applied to select features in previous studies (Yeo, Cameron et al. 1995, Kokaly and Clark 1999, Chou, Tiu et al. 2001, Kim and Kim 2004, Chen and Lin 2006, Díaz-Uriarte and De Andres 2006, Saeys, Inza et al. 2007, Ma and Huang 2008) and demonstrated good performance in feature selection. We claim a feature as an important feature if at least two of the four methods consider this feature important. By applying the four methods to the training data, we selected 13 important features (Section 3.1).

3.2.2. 4 TarPmiR, a random-forest based approach for miRNA target site prediction

With the 13 selected features, we developed a random-forest based approach called TarPmiR for miRNA target site prediction. We chose the random forests method because we applied the above four approaches to the aforementioned training and testing datasets and found that random forests gave the best performance (Section 3.2).

TarPmiR predicts miRNA target sites in three steps with the input of a set of miRNAs and a set of mRNAs. First, TarPmiR generates candidate target sites based on seed match or minimal folding energy (Enright, John et al. 2004, Grimson, Farh et al. 2007, Yousef, Jung et al. 2007). For a given miRNA, TarPmiR scans an mRNA sequence with the seed region of the miRNA (positions 2-7) to find perfect seed-matching sites. These sites are defined as the first set of candidate target sites. In addition, TarPmiR applies RNA-duplex from the Vienna RNA package (Hofacker 2003) to obtain

the top target sites with the lowest folding energy. These energy-based sites are defined as the second set of candidate target sites. The combination of seed match and folding energy helps TarPmiR to pick up almost all true target sites from the beginning. Second, for each candidate target sites, TarPmiR calculates the values of the 13 selected features. Finally, TarPmiR applies the trained random-forest based predictor to predict target sites. The output of the random-forest model is the predicted probability that a candidate target site is a true target site. We have compared nine probability cutoffs to define target sites using the F2 score, since we put more emphasis on the recall than the precision. The cutoffs 0.5 and 0.6 have almost the similar F2 scores, while the cutoff 0.5 has the largest recall. Therefore, we used 0.5 for the following analyses. We provide a parameter –p in TarPmiR, users can choose other cutoffs based on their own needs.

3.2.2. 5 Comparisons with other methods

We compared TarPmiR with the following methods: targetScan (8,13), miRanda (9), and miRmap (Vejnar and Zdobnov 2012, Vejnar, Blum et al. 2013). The targetScan and miRanda are two of the most widely used miRNA target prediction tools. We used the following commands to run them: perl targetscan.pl <miRNA.> <mRNA> <targetscan_out>; perl targetscan_60_context_scores.pl <miRNA> <mRNA> <targetscan_out> <targetscan_context_score_out>; and miranda <miRNA> <mRNA> -sc 120 –en 1. MiRmap is a recently developed tool, which takes high throughput sequencing data as input to predict miRNA target sites. MiRmap provides a python library and users can write a script to output miRmap predictions with the functions in the library. We used similar parameters as in (Vejnar and Zdobnov 2012) when running miRmap.

3.2.3 Results

3.2.3.1 All But One Conventional Features and Seven New Features Were Selected by Different Approaches.

We applied four approaches to select important features from the 18 potential features. Each approach selected a similar but slightly different subset of features. By defining features selected by at least two approaches as important features, we discovered 13 important features (Figure 3-4). They are: 1) Folding energy; 2) Seed match; 3) Accessibility; 4) AU Content, 5) Stem conservation; 6) Flanking conservation; 7) m/e motif; 8) The total number of paired positions; 9) The length of the target mRNA region; 10) The length of the largest consecutive pairings; 11) The position of the largest consecutive pairings relative to the 5' end of miRNA; 12) The number of paired positions at the miRNA 3' end. Recall miRNA 3' end meant the last 7 positions of a miRNA; 13) The difference between the number of paired positions in the seed region and that in the miRNA 3' end.

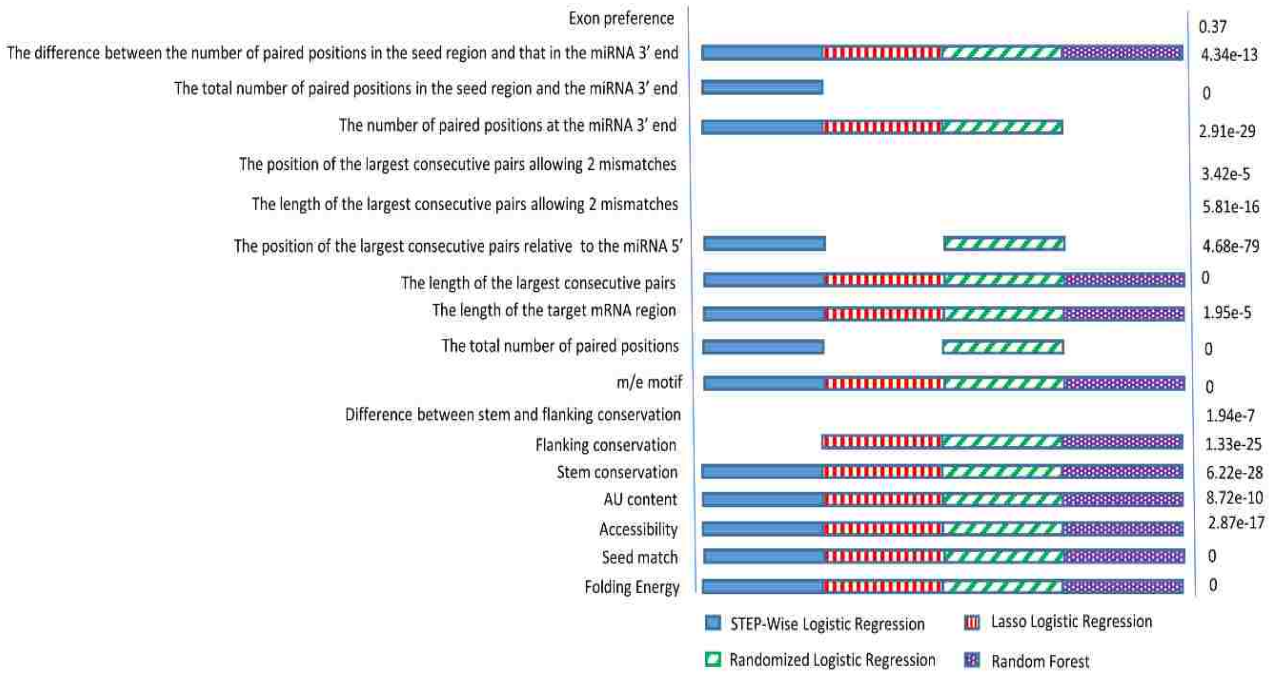

Figure 3-4 Feature selection

An interesting observation from Figure 3-4 was the removal of one and only one conventional feature, the difference between stem and flanking conservation. This feature was used in previous studies (Pollard, Hubisz et al. 2010, Helwak, Kudla et al. 2013). The removal of this feature may be explained by the fact that most positive target sites from CLASH experiments were from coding regions and there was not much difference in terms of conservation between the seed regions and the flanking regions of target sites in coding regions. Because true target sites were functional and conserved, two features related to the conservation in miRNA-mRNA stem regions and in flanking regions around the stems, respectively, were selected.

In addition to the six selected conventional features (folding energy, seed-matching, accessibility, AU content), four new features were selected by all four approaches (Figure 3-4). These features were the m/e motif, the length of the target site, the length of the largest consecutive pairings, and the difference between the number of paired positions in the seed region and that in the miRNA 3' end. The inclusion of the m/e motif implied that there existed preferred matching positions shared by all miRNAs. The length of the target site was selected, showing the importance of the binding preference of miRNAs to mRNA regions with specific lengths. The length of the largest consecutive pairing positions mattered, which extended the concept of seed match, as seed match was just a simple case with a long consecutive pairing positions. The difference between the number of paired positions in the seed region and that in the miRNA 3' end also suggested that the seed match may be unimportant, given a high-quality 3' end region matching. This also supported the idea that a long consecutive matching region is critical for functional miRNA target sites.

We further investigated the importance of the 13 selected features by the rank-sum test (Mann and Whitney 1947) (Figure 3-4). In brief, for each selected feature, we calculated its value for all positive target sites and for their corresponding negative target sites. We then compared the two groups of numbers by the rank-sum test. The numbers on the right side of Figure 3-4 showed the p-values of the corresponding features. All 13 selected features had a significant difference between the positive target sites and negative sites (p-value<1.95e-5). Some significant features based on the rank-sum test were not selected by the four machine learning methods, which may be due to the fact that the contribution from the combination of the selected 13 features can already replace that of these removed features. In fact, we calculated the correlation between every pair of the 18 features and found that the discarded significant features correlate well with certain important features.

3.2.3.2 TarPmiR Had a >55% Recall and a >19.1% Precision.

With the 13 selected features, we developed the TarPmiR method to predict miRNA target sites in the entire regions of mRNAs. TarPmiR applied the random-forest based approach for target site prediction. It applied the random forests approach instead of the other three approaches because when tested on five testing datasets, the random forests based approach always gave better recalls and precisions (Table 3.3).

Table 3.3 Recall and precision of different methods on five testing datasets

	Lasso Logistic		Randomized Logistic		STEP-Wise Logistic		Random Forest		TarPmiR	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
T1	0.8549	0.7765	0.8539	0.7785	0.8559	0.7795	0.8740	0.8283	0.5514	0.1905
T2	0.8736	0.7713	0.8746	0.7730	0.8751	0.7736	0.8921	0.8296	0.5227	0.1626
T3	0.8315	0.7626	0.8319	0.7898	0.8320	0.7904	0.8686	0.8253	0.5303	0.1661
T4	0.836	0.7871	0.8411	0.7903	0.838	0.7894	0.8776	0.8266	0.5507	0.1902
T5	0.8856	0.7639	0.8878	0.7662	0.8895	0.7649	0.8989	0.8173	0.5583	0.1909

To investigate the recall and precision of TarPmiR, we tested it on the five testing datasets described in Material and Methods. The precision and recall of TarPmiR in each set of test data were shown in Table 3.3. Since the TarPmiR predictors built on each of the five training datasets had similar precision and recall, we chose the first TarPmiR predictor in our developed tool and in the following analyses. TarPmiR had a 55.1% recall and a 19.1% precision, which were higher than the recall and precision of existing methods reviewed in (Reczko, Maragkakis et al. 2011). Note that TarPmiR had a much smaller precision and recall than the above four methods (columns 2 to 5 in Table 3.3), because it predicted target sites from the entire mRNA sequences instead of the 8514 sites that were not used for training.

3.2.3.3 TarPmiR Predicted the Majority of True Target Sites in Independent Datasets.

To investigate whether TarPmiR was able to predict true target sites in non-CLASH datasets, we applied it to two PAR-CLIP datasets in the

HEK293 cell (Material and Methods). There were 16041 "true" target sites in 10023 mRNAs from the first dataset (dataset I). Moreover, the reads of the top 60 miRNAs and top 120 miRNAs accounted for more than 90% and 99%, respectively, of the total PAR-CLIP reads in this dataset. By inputting 60 miRNAs and 10023 mRNAs, TarPmiR predicted 240605 target sites, which included 74.2% of true target sites (Table 3.4). Similarly, by inputting 120 miRNAs and 10023 mRNAs, TarPmiR predicted 481135 target sites, which included 86.3% of true target sites (Table 3.4). The percentages of correctly predicted true target sites should be considered underestimated, as a portion of true target sites may not be target sites of the 60 or 120 miRNAs. By considering the 16041 "true" target sites as all target sites in these mRNAs, we found that TarPmiR had a >74% recall in this dataset (Table 3.4). For the second PAR-CLIP dataset (dataset II), there were 43251 "true" target sites in 17794 mRNAs. Because the cell was the same as that in the first PAR-CLIP dataset, we assumed that mainly 60 or 120 miRNAs related to these target sites. Similarly, we found that TarPmiR was able to identify 79.3% and 89.8% of "true" target sites, when inputting 60 miRNAs and 120 miRNAs, respectively, together with the 17794 mRNAs (Table 3.4).

Table 3.4 Comparison of four methods on independent datasets

Data set	# of miRNAs input	Performance measurement	TarPmiR	miRanda	TargetScan	miRmap
I	60	# of predictions	240605	246311	219304	504447
		% of correct predictions	11904/16041=74.2%	7061/16041=44.0%	6248/16041=39.0%	7121/16041 =44.4%
		Recall	0.742	0.440	0.390	0.444
		Precision	0.0495	0.0287	0.0285	0.014
	120	# of predictions	481135	476827	461280	906654
		% of correct predictions	13846/16041=86.3%	9683/16041=60.4%	8969/16041=55.9%	10342/16041=64.5%
		Recall	0.863	0.604	0.559	0.645
		Precision	0.0288	0.0203	0.0194	0.0114
II	60	# of predictions	469752	453880	437791	971238
		% of correct predictions	34301/43251 =79.3%	20378/43251 =47.1%	17556/43251 =40.6%	20543/43251 =47.5%
		Recall	0.793	0.471	0.406	0.475
		Precision	0.0730	0.0449	0.0401	0.0211
	120	# of predictions	961112	902611	922373	1952258
		% of correct predictions	38821/43251= 89.8%	23762/43251= 54.9%	24578/43251= 56.8%	25667/43251= 59.3%
		Recall	0.898	0.549	0.568	0.593
		Precision	0.0403	0.0263	0.0266	0.0131
III	119	# of predictions	285491	439485	875442	341773
		% of correct predictions	10766/11080=97.2%	9069/11080=81.8%	10084/11080=91.0%	7840/11080=70.8%
		Recall	0.972	0.818	0.910	0.708
		Precision	0.0377	0.0206	0.0115	0.0229
IV	50	# of predicted interactions	102324	87462	69184	113211
		% of correct predictions	5515/15386=35.8%	4335/15386=28.2%	3821/15386=24.8%	4531/15386=29.4%
		Recall	0.358	0.282	0.248	0.294
		Precision	0.0539	0.0496	0.0552	0.0400
	100	# of predicted interactions	412149	337863	286667	413213
		% of correct predictions	8127/15562=52.2%	6442/15562=41.4%	5644/15562=36.3%	5323/15562=34.2%
		Recall	0.522	0.414	0.363	0.342
		Precision	0.0197	0.0191	0.0197	0.0129

The above analyses demonstrated the successful performance of TarPmiR in the human dataset in the same cell type. It was unclear how well TarPmiR performed in other species and in other cell types. We thus applied TarPmiR to a third independent dataset, the mouse HITS-CLIP dataset in the cortex cell (dataset III). There were 119 potential miRNAs and 2953 mRNAs involved in a total of 11080 target sites. With the input of these 119 miRNAs and 2953 mRNAs, TarPmiR predicted 285491 target sites in total. There were 10766 of the 11080 (97.2%) target sites predicted by TarPmiR (Table 3.4).

In addition to the above analyses on the crosslinking based data, we tested TarPmiR using the annotated miRNA-mRNA interactions in TarBase 6.0 (dataset IV) (Table 3.4). For the top 50 miRNAs and the corresponding 9823 target mRNAs, TarPmiR predicted 35.8% of true target mRNAs (Methods). For the top 100 miRNAs and the corresponding 9869 target mRNAs, TarPmiR predicted 52.2% of true target mRNAs (Table 3.4) (Methods).

3.2.3.4 TarPmiR Showed Superior Performance to Existing Approaches.

We compared TarPmiR with two widely used tools miRanda (9) , targetScan (8,13) and a recently published tool, miRmap (Vejnar and Zdobnov 2012, Vejnar, Blum et al. 2013). The comparison was made on the CLASH dataset, the three independent datasets, and the two databases described above. Overall, TarPmiR with the default cutoff 0.5 had a much higher recall and precision than the three existing methods on the CLASH dataset (Table 3.5). For instance, TarPmiR had a recall of 55.1%, which was at least 10% higher than other approaches. TarPmiR had a precision of 19.1%, which was at least 9% higher than other approaches.

Table 3.5 Comparison of different methods on the CLASH dataset

Method	TP	FN	FP	Recall TP/(TP+FN)	Precision TP/(TP+FP)	F1-score
TarPmiR	4695	3819	283462	0.551	0.191	0.284
miRanda	3852	4662	285708	0.452	0.069	0.120
TargetScan	1164	7350	213885	0.136	0.101	0.116
Mirmap	1821	6693	297610	0.214	0.056	0.089

On the three independent datasets, we compared TarPmiR with the other three methods (Table 3.4). Overall, TarPmiR had a similar or much smaller number of predicted target sites, while it had much more known miRNA target sites predicted in each dataset. By assuming the CCRs

from PAR-CLIP and target sites from HITS-CLIP were the only true miRNA target sites in the corresponding mRNAs in the corresponding datasets, we found that TarPmiR had a recall at least 6.2% higher than other methods, and a precision at least 0.85% higher than other methods. Note that the performance of all four methods was relatively high in the mouse dataset than other independent datasets, because miRNA-mRNA interactions in this dataset were mainly inferred and majorly based on seed regions (Chi, Zang et al. 2009).

For the known miRNA-mRNA interactions in TarBase 6.0, we also compared TarPmiR with other three methods (Table 3.4). TarPmiR had a similar or slightly larger number of predicted interactions, while it predicted much more known miRNA-mRNA interactions. Similar to the results on crosslinking based datasets, TarPmiR had a much higher recall and a higher precision than other methods.

We also compared the running speed of the four methods. Because TarPmiR was a machine learning based method and it calculated more features, it was much slower than miRanda and TargetScan. The running speed was similar to that of Mirmap, which was also a machine learning based method. It was worth pointing out that, although TarPmiR was relatively slow, its speed was reasonable. For instance, it took TarPmiR about 7940 CPU seconds to predict target sites of 20 miRNAs in 400 mRNA sequences, on average each 2000 nt long.

3.2.4 Discussion

In this study, we identified seven new features together with six conventional features of miRNA target sites. Based on these 13 selected features, we developed a new approach called TarPmiR to predict miRNA target sites. We tested TarPmiR on a human CLASH dataset, two human PAR-CLIP datasets, a mouse HITS-CLIP dataset, and a general dataset from TarBase 6.0, and showed that TarPmiR performed at least the same or better than three existing approaches. TarPmiR is freely available at http://hulab.ucf.edu/research/projects/miRNA/TarPmiR/ .

Not all new features were completely new. We claimed some features as new because they were not used by most of the existing tools, such as miRanda (Enright, John et al. 2004), TargetScan (Grimson, Farh et al. 2007, Friedman, Farh et al. 2009), DIANA-microT-CDS (Maragkakis, Reczko et al. 2009, Paraskevopoulou, Georgakilas et al. 2013), rna22-gui (Loher and Rigoutsos 2012), TargetMiner (Bandyopadhyay and Mitra 2009), PITA (Kertesz, Iovino et al. 2007), RNAhybrid (Krüger and Rehmsmeier 2006), etc. However, several new features were mentioned in previous studies directly or indirectly. For instance, Thomason et al stated that "some validated miRNA target sites do not have a complete seed match but instead exhibit 11–12 continuous base pairs in the central region of the miRNA" (Thomson, Bracken et al. 2011). We observed similar target sites in the CLASH dataset and proposed the feature "The length and position of the longest consecutive pairs".

The selected new features significantly improved the prediction accuracy of TarPmiR. To show the contribution of the new features to the accuracy of TarPmiR, we removed the seven new features and retrained random forests in TarPmiR. Compared with the original TarPmiR with 13 features, the recall and precision of the modified TarPmiR dropped 8.6% and 9.7%, respectively. We also compared the predicted true target sites by different approaches. TarPmiR had the largest number of predicted true sites shared by other tools. However, the percentage of shared true target sites predicted by TarPmiR was lower than that of other tools, suggesting that TarPmiR complements existing tools by predicting sites that cannot be predicted by other tools. In fact, there are 2090 'non-seed-matching' sites in the first CLASH test dataset. TarPmiR was able to identify 1585 (75.8%) of those sites. On the other hand, miRanda and TargetScan were only able to predict 173 (8.28%) and 34 (1.6%) sites, respectively. This also suggested that the traditional tools like TargetScan and miRanda almost cannot predict non-seed-matching binding sites.

It is also worth mentioning that CLASH experiments may pick up direct and indirect miRNA target sites. The Argonaut proteins are guided by miRNAs to bind mRNAs, which is referred to as miRNA-dependent recruitment and results in direct miRNA target sites. There is also a miRNA-independent Argonaut protein recruitment mechanism, in which Argonaut proteins are recruited to target mRNAs by protein-protein interaction with RNA-binding proteins and thus miRNAs do not interact with the mRNAs directly (Meister 2013). In the future, one may want to distinguish these two types of target sites from the CLASH experiments before training

predictors for target site prediction. In this way, we may also obtain better features and improve the prediction accuracy.

Because of the existence of indirect target sites in CLASH data, the recall of TarPmiR on the CLASH testing datasets may be underestimated. In fact, TarPmiR had a much higher recall on the three independent human and mouse datasets, suggesting that TarPmiR may have a recall larger than 74%. On the other hand, TarPmiR had a much lower precision on the independent datasets, which may be underestimated as well. This was because we treated all segments other than the CCRs or identified miRNA target sites in these independent datasets as true negative target sites, which may not be the case.

By the time of this study, only one CLASH dataset was publicly available (Helwak, Kudla et al. 2013). This human CLASH dataset was used to train TarPmiR. We applied TarPmiR to human and mouse datasets and demonstrated that it works well on these datasets. In the future, with more CLASH datasets available, more important miRNA target site features including tissue-specific features may be discovered and the accuracy of TarPmiR, especially its precision, may be further improved.

CHAPTER 4: CONCLUSION AND FUTURE WORK

4.1 Conclusion

This dissertation mainly focuses on the gene regulation studies at both the transcriptional and post-transcriptional levels. At the transcription level, the studies are mainly on the TF binding. Systematic discovery of TF binding sites (TFBSs) and binding motifs is crucial for the study of gene transcriptional regulation (Birney, Stamatoyannopoulos et al. 2007). The binding of TFBSs by TFs can activate or repress the transcription of genes near the TFBSs, thus can modulate gene expression. In this dissertation, we proposed 3 TF binding prediction tools: MERCED, SIOMICS, and SIOMICS2. These tools outperformed the existing TF binding prediction tools such as DREME in terms of prediction accuracy and running efficiency. They are very helpful for improving the understanding of transcriptional regulation mechanism, especially on TF binding.

MERCED: Genome-wide identification of CREs in *C. reinhardtii* genome is critical to further study gene regulation and molecular functions. We thus performed the first large-scale CRE prediction in *C. reinhardtii*. Different from available CRE prediction methods, our approach considers the species-divergence-time and the nucleotide content rather than depends on MSA and mismatch-number-counting to determine whether DNA segments are conserved CREs and motifs. According to sequence permutation test, the developed method has very low false discovery rate. Compared with alignment and mismatch-number-counting-based approaches, the developed method is more efficient in

filtering divergent segments while keeping conserved segments that have quite a few mismatches compared with their counterpart segments.

SIOMICS: We developed a novel approach SIOMICS to systematically discover motifs and TFBSs from ChIP-seq data. Different from available methods, SIOMICS does not depend on limited information of known motifs and simultaneously considers multiple motifs. Tested on experimental and simulated data, we shown that SIOMICS identifies motifs of more known cofactors and identifies more shared motifs in the experimental data. At the same time, SIOMICS has a low false positive rate when tested in the simulated data. In addition, we shown SIOMICS is as fast as other methods especially when the ChIP-seq datasets are large. SIOMICS is thus a useful alternative method for motif discovery.

SIOMICS2: We developed a useful tool set SIOMICS for systematic discovery of motifs in ChIP-seq datasets. Tested on 13 real ChIP-seq datasets and 13 random datasets, we showed that SIOMICS predicted motifs of known cofactors in each real dataset while no motif in any random dataset. In this extension version of the SIOMICS tool – SIOMICS_Extension, we modified the motif length of a motif based on its predicted TFBSs from all motif modules containing this motif. The developed SIOMICS tool set (Original SIOMICS and SIOMICS_Extension) together with its manual, test datasets, and others is freely available at http://www.cs.ucf.edu/~xiaoman/SIOMICS/SIOMICS.html.

We will continue to keep the updated versions on this site to help researchers fully utilize the potential of the ChIP-seq data.

At the post-transcription level, the studies are mainly on the microRNA targeting. MicroRNAs is one of the most abundant gene regulatory molecules in animals and plants. In human, there are over 1000 miRNAs identified. MicroRNA are reported to regulate more than 60% of protein-coding genes in human the miRNAs and they are also involved in diverse regulatory pathways, as well as in disease development, progression, prognosis, and treatment. Even through the microRNAs are very important, our understanding of their regulatory functions is still very limited. As a result, the cellular functions and pathways, which are affected by microRNA in various cancers, are still unknown in most cases. Therefore, one of the major challenges of current miRNA study is identifying the regulatory targets of miRNAs. In this studies, we proposed several miRNA binding target prediction tools, which show advantages or superior performance compared with existing tools.

miRModule: We studied miRNA modules based on experimentally determined miRNA target sites. We predicted 181 miRNA modules and 306 potential miRNA modules. We demonstrated that miRNA modules preferred to bind weak sites and favored a combination of all unconventional sites. We also observed that miRNA modules preferred to bind in CDSs and favored the first and the last exons. We confirmed that more than 70% of miRNA modules bound sites within specific ranges, with enrichment in two previously known ranges. However, many more adjacent sites bound by miRNA modules were >130 nucleotides apart. We further showed that unconventional target sites

of miRNA modules were often within shorter distances than other combinations of target sites. Our study shed new light on miRNA binding.

TarPmiR: In this study, we identified seven new features together with six conventional features of miRNA target sites. Based on these 13 selected features, we developed a new approach called TarPmiR to predict miRNA target sites. We tested TarPmiR on a human CLASH dataset, two human PAR-CLIP datasets, a mouse HITS-CLIP dataset, and a general dataset from TarBase 6.0, and showed that TarPmiR performed at least the same or better than three existing approaches. TarPmiR is freely available at http://hulab.ucf.edu/research/projects/miRNA/TarPmiR/ .

4.2 Future Work

4.2.1 SIOMICS3

There are approximately 2600 proteins in the human genome that contain DNA-binding domains, most of which are presumed to function as transcription factors (Babu, Luscombe et al. 2004). There are about 1500 transcription factors (TFs) that play an important role in transcriptional regulation by binding to the genome alone or in complexes (Vaquerizas, Kummerfeld et al. 2009). Chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing (ChIP-seq) is the current method of choice to study the genomic binding locations of transcription factors, as well as the localization of epigenetic regulatory marks. A typical TF ChIP-Seq experiment generally yields tens of thousands of predicted binding locations (TF ChIP-seq peaks).

Plenty of specialized software tools have been recently developed for motif discovery of TF ChIP-seq peaks. **One common bottleneck** of many exiting tool is that the underlying algorithm was originally designed for motif discovery from a small set of co-regulated promoters, and therefore they can hardly handle the large scale TF ChIP-seq data. Those tools always truncate the TF ChIP-seq peaks or reduce the number of the peaks to circumvent the limitation. **A second limitation is** that most algorithms are traditionally monad-based and therefore they can only predict monad motifs. All the dyad motifs are completely ignored. Dyad is pairs of short oligonucleotide (3–4 bp), separated by a spacing of fixed width but variable content (e.g.,

CTA…TGG) (Defrance, Sand et al. 2008), while the monad motifs are short contiguous patterns. **A third limitation** is that most algorithms do not consider co-working of different TFs. They did not provide the information of which TFs always work together. (We name it motif modules). **A fourth limitation** is that lots of current ChIP-Seq motif finding tools are running slow especially when the number of TF ChIP-seq peaks is large.

Facing those limitations, we are going to improve the existing SIOMICS pipeline to fulfill the following practical needs. First of all, it can handle large scale TF ChIP-seq peaks. There is no limitation on number or width of input peaks. Second, it can predict both the monad and dyad motif at the same time. Third, it can automatically predict motif modules so that you might have a sight on which TFs are working together. Last but not least, the time efficiency should be comparable to other widely used tools especially when input data set is large.

4.2.2 miRHMM

microRNAs (miRNAs) are ~22 nt endogenous RNAs involved in the regulation of gene expression(Ambros 2001, Bartel 2004). The regulatory functions of microRNAs are carried out through the RNA-induced silencing complex (RISC). The microRNAs guide the binding of the RISC through the base-paring to the target mRNA and thereby negatively regulate its expression. MicroRNAs is one of the most abundant gene regulatory molecules in animals and plants. In human, there are over 1000 miRNAs identified (Griffiths-Jones, Grocock et al. 2006) and each of them has the potential to bind to hundreds of target genes. More than 60% of protein-coding genes appear to be regulated by microRNAs (Friedman, Farh et al. 2009). Besides, the miRNAs are involved in diverse regulatory pathways (Chang, Johnston et al. 2004, Bilen, Liu et al. 2006, Bilen, Liu et al. 2006), as well as in disease development, progression, prognosis, and treatment(Yang, Coukos et al. 2008, Yu, Chen et al. 2008). For example, human miR-155 has been reported to mediate T cell dependent anti-body response(Rodriguez, Vigorito et al. 2007, Thai, Calado et al. 2007) and it was also involved in many cancers like lymphomas, breast cancer, lung, and colon cancer.(Metzler, Wilda et al. 2004, Kluiver, Poppema et al. 2005, Yanaihara, Caplen et al. 2006).

Even through the microRNAs are very important, our understanding of the regulatory functions of them is still very limited. As a result, the cellular functions and pathways, which are affected

by microRNA in various cancers, are still unknown in most cases. Therefore, one of the major challenges of current miRNA study is identifying the regulatory targets of miRNAs. Because of a huge number of microRNAs and their potential targets, a mere experiment prediction design is very laborious and expensive. Computational target prediction methods together with the high-throughput experiments would be a great choice for identifying miRNA targets.

A very large number of computational miRNA target prediction methods has been published in the past decade. They were proposed based on different sets of miRNA binding features. TargetScan (Lewis, Shih et al. 2003, Lewis, Burge et al. 2005, Grimson, Farh et al. 2007) was first proposed in 2003 and it's still one of most popular miRNA target prediction algorithms nowadays. Those early proposed miRNA target prediction tools, limited by the availability of relevant experiment data, are mainly based on these traditional features, including seed, energy, conservation, accessibility etc. (Peterson, Thompson et al. 2014). These 'traditional feature' based algorithms include TargetScan (Lewis, Shih et al. 2003, Lewis, Burge et al. 2005, Grimson, Farh et al. 2007), miRanda(Enright, John et al. 2004, John, Enright et al. 2004, Betel, Wilson et al. 2008), PITA(Kertesz, Iovino et al. 2007), RNAhybrid (Krüger and Rehmsmeier 2006), etc. Recently, many new miRNA target prediction algorithms were proposed as the increasing availability of the relevant miRNA target experiments, such as CLIP-seq (Ule, Jensen et al. 2003), PAR-CLIP (Hafner, Landthaler et al. 2010), CLASH (Helwak, Kudla et al. 2013), etc. These new miRNA target prediction algorithms were based on the important new features learned from the experiment data using statistical/ machine learning models. These 'new data'

driven algorithms include microT-CDS (Reczko, Maragkakis et al. 2012, Paraskevopoulou, Georgakilas et al. 2013), MiRmap (Vejnar and Zdobnov 2012), MirTarget (Wang and El Naqa 2008), PicTar (Krek, Grün et al. 2005), etc.

Most existing miRNA target prediction tools, both tradition feature based and new data based, are mainly considering the miRNA target prediction in a site-specific manner. These tools are only using the binding features between each pair of miRNA and mRNA to predict. However, the efficiency of miRNA-mediated regulation can be affected by multiple system-wide factors (not binding site-specific) (Arvey, Larsson et al. 2010, Sumazin, Yang et al. 2011), such as miRNA/mRNA expression level and combinatorial binding of multiple microRNAs. None of the traditional feature-based methods (e.g. TargetScan, miRanda) has taken the system-wide factors into consideration. Some of the new data-driven tools tried to make use of the system-wide factors but they are also very limited. For example, PicTar (Krek, Grün et al. 2005) is a computational method to identify the common target for miRNAs, but it does not even take the miRNA expression in determining the relative binding. To fill this gap, we need to develop a miRNA binding prediction tools, which can take the system-wide factors such as miRNA expression into consideration.

LIST OF REFERENCES

Abe, H., et al. (2003). "Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling." Plant Cell **15**(1): 63-78.

Aldous, D. (1989). Probability Approximations via the Poisson Clumping Heuristic, Springer-Verlag.

Allocco, D. J., et al. (2004). "Quantifying the relationship between co-expression, co-regulation and gene function." BMC Bioinformatics **5**: 18.

Altman, R. B. and S. Raychaudhuri (2001). "Whole-genome expression analysis: challenges beyond clustering." Curr Opin Struct Biol **11**(3): 340-347.

Altschul, S. F., et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.

Ambros, V. (2001). "microRNAs: tiny regulators with great potential." Cell **107**(7): 823-826.

Ambros, V. (2004). "The functions of animal microRNAs." Nature **431**(7006): 350-355.

Amr, A. G., et al. (2006). "Anticancer activities of some newly synthesized pyridine, pyrane, and pyrimidine derivatives." Bioorg Med Chem **14**(16): 5481-5488.

Arnone, M. I. and E. H. Davidson (1997). "The hardwiring of development: organization and function of genomic regulatory systems." Development **124**(10): 1851-1864.

Arvey, A., et al. (2010). "Target mRNA abundance dilutes microRNA and siRNA activity." Molecular systems biology **6**(1): 363.

Ashburner, M., et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Axtell, M. J., et al. (2011). "Vive la difference: biogenesis and evolution of microRNAs in plants and animals." Genome Biol **12**(4): 221.

Babu, M. M., et al. (2004). "Structure and evolution of transcriptional regulatory networks." Current opinion in structural biology **14**(3): 283-291.

Bailey, T. L. (2011). "DREME: motif discovery in transcription factor ChIP-seq data." Bioinformatics **27**(12): 1653-1659.

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proc Int Conf Intell Syst Mol Biol **2**: 28-36.

Bailey, T. L., et al. (2006). "MEME: discovering and analyzing DNA and protein sequence motifs." Nucleic acids research **34**(suppl 2): W369-W373.

Bandyopadhyay, S. and R. Mitra (2009). "TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples." Bioinformatics **25**(20): 2625-2631.

Bartel, D. P. (2004). "MicroRNAs: genomics, biogenesis, mechanism, and function." Cell **116**(2): 281-297.

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

Beckmann, J., et al. (2009). "Improvement of light to biomass conversion by de-regulation of light-harvesting protein translation in Chlamydomonas reinhardtii." J Biotechnol **142**(1): 70-77.

Betel, D., et al. (2008). "The microRNA. org resource: targets and expression." Nucleic acids research **36**(suppl 1): D149-D153.

Bilen, J., et al. (2006). "A new role for microRNA pathways: modulation of degeneration induced by pathogenic human disease proteins." Cell Cycle **5**(24): 2835-2838.

Bilen, J., et al. (2006). "MicroRNA pathways modulate polyglutamine-induced neurodegeneration." Molecular cell **24**(1): 157-163.

Birney, E., et al. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Bisova, K., et al. (2005). "Genome-wide annotation and expression profiling of cell cycle regulatory genes in Chlamydomonas reinhardtii." Plant Physiol **137**(2): 475-491.

Blanchette, M., et al. (2006). "Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression." Genome Res **16**(5): 656-668.

Blanchette, M. and M. Tompa (2002). "Discovery of regulatory elements by a computational method for phylogenetic footprinting." Genome Res **12**(5): 739-748.

Bohne, F. and H. Linden (2002). "Regulation of carotenoid biosynthesis genes in response to light in Chlamydomonas reinhardtii." Biochim Biophys Acta **1579**(1): 26-34.

Boyle, E. I., et al. (2004). "GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes." Bioinformatics **20**(18): 3710-3715.

Brennecke, J., et al. (2005). "Principles of microRNA-target recognition." PLoS Biol **3**(3): e85.

Bruce, W. B., et al. (1991). "A negatively acting DNA sequence element mediates phytochrome-directed repression of phyA gene transcription." Embo J **10**(10): 3015-3024.

Bruce, W. B. and P. H. Quail (1990). "cis-acting elements involved in photoregulation of an oat phytochrome promoter in rice." Plant Cell **2**(11): 1081-1089.

Bryan, K., et al. (2013). "Discovery and visualization of miRNA-mRNA functional modules within integrated data using bicluster analysis." Nucleic Acids Res **42**(3): e17.

Busk, P. K. and M. Pages (1997). "Protein binding to the abscisic acid-responsive element is independent of VIVIPAROUS1 in vivo." Plant Cell **9**(12): 2261-2270.

Busk, P. K. and M. Pages (1998). "Regulation of abscisic acid-induced transcription." Plant Mol Biol **37**(3): 425-435.

Cai, X., et al. (2010). "Systematic identification of conserved motif modules in the human genome." BMC Genomics **11**: 567.

Cardol, P., et al. (2005). "The mitochondrial oxidative phosphorylation proteome of Chlamydomonas reinhardtii deduced from the Genome Sequencing Project." Plant Physiol **137**(2): 447-459.

Castresana, C., et al. (1988). "Both positive and negative regulatory elements mediate expression of a photoregulated CAB gene from Nicotiana plumbaginifolia." Embo J **7**(7): 1929-1936.

Castruita, M., et al. (2011). "Systems biology approach in Chlamydomonas reveals connections between copper nutrition and multiple metabolic steps." Plant Cell **23**(4): 1273-1292.

Chang, H. Y., et al. (2004). "Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds." PLoS Biol **2**(2): E7.

Chang, S., et al. (2004). "MicroRNAs act sequentially and asymmetrically to control chemosensory laterality in the nematode." Nature **430**(7001): 785-789.

Chatr-Aryamontri, A., et al. (2013). "The BioGRID interaction database: 2013 update." Nucleic Acids Res **41**(Database issue): D816-823.

Chen, X., et al. (2008). "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells." Cell **133**(6): 1106-1117.

Chen, Y.-W. and C.-J. Lin (2006). Combining SVMs with various feature selection strategies. Feature extraction, Springer**:** 315-324.

Chi, S. W., et al. (2009). "Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps." Nature **460**(7254): 479-486.

Chou, C.-H., et al. (2013). "A computational approach for identifying microRNA-target interactions using high-throughput CLIP and PAR-CLIP sequencing." BMC genomics **14**(Suppl 1): S2.

Chou, Y.-H., et al. (2001). "Stepwise logistic regression analysis of tumor contour features for breast ultrasound diagnosis." Ultrasound in medicine & biology **27**(11): 1493-1498.

Defrance, M., et al. (2008). "Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences." Nature protocols **3**(10): 1589-1603.

Díaz-Uriarte, R. and S. A. De Andres (2006). "Gene selection and classification of microarray data using random forest." BMC bioinformatics **7**(1): 3.

Didiano, D. and O. Hobert (2006). "Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions." Nature structural & molecular biology **13**(9): 849-851.

Ding, J., et al. (2013). "Chipmodule: systematic discovery of transcription factors and their cofactors from chip-seq data." Pac Symp Biocomput: 320-331.

Ding, J., et al. (2015). "Systematic discovery of cofactor motifs from ChIP-seq data by SIOMICS." Methods **79**: 47-51.

Ding, J., et al. (2012). "Thousands of cis-regulatory sequence combinations are shared by Arabidopsis and poplar." Plant Physiol **158**(1): 145-155.

Ding, J., et al. (2013). "SIOMICS: a novel approach for systematic identification of motifs in ChIP-seq data." Nucleic Acids Res **42**(5): e35.

Ding, J., et al. (2014). "SIOMICS: a novel approach for systematic identification of motifs in ChIP-seq data." Nucleic Acids Res **42**(5): e35.

Ding, J., et al. (2012). "Systematic prediction of cis-regulatory elements in the Chlamydomonas reinhardtii genome using comparative genomics." Plant Physiol **160**(2): 613-623.

Ding, J., et al. (2015). "MicroRNA modules prefer to bind weak and unconventional target sites." Bioinformatics **31**(9): 1366-1374.

Doench, J. G. and P. A. Sharp (2004). "Specificity of microRNA target selection in translational repression." Genes Dev **18**(5): 504-511.

Donohoe, M. E., et al. (2009). "The pluripotency factor Oct4 interacts with Ctcf and also controls X-chromosome pairing and counting." Nature **460**(7251): 128-132.

Eberhard, S., et al. (2006). "Generation of an oligonucleotide array for analysis of gene expression in Chlamydomonas reinhardtii." Curr Genet **49**(2): 106-124.

Edgar, R., et al. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." Nucleic Acids Res **30**(1): 207-210.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-1797.

Elemento, O. and S. Tavazoie (2005). "Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach." Genome Biol **6**(2): R18.

Elemento, O. and S. Tavazoie (2007). "Fastcompare: a nonalignment approach for genome-scale discovery of DNA and mRNA regulatory elements using network-level conservation." Methods Mol Biol **395**: 349-366.

Enright, A. J., et al. (2004). "MicroRNA targets in Drosophila." Genome biology **5**(1): R1-R1.

Fauteux, F., et al. (2008). "Seeder: discriminative seeding DNA motif discovery." Bioinformatics **24**(20): 2303-2307.

Fischer, B. B., et al. (2009). "Function and regulation of the glutathione peroxidase homologous gene GPXH/GPX5 in Chlamydomonas reinhardtii." Plant Mol Biol **71**(6): 569-583.

Fischer, B. B., et al. (2012). "SINGLET OXYGEN RESISTANT 1 links reactive electrophile signaling to singlet oxygen acclimation in Chlamydomonas reinhardtii." Proc Natl Acad Sci U S A **109**(20): E1302-1311.

Fogel, B. L., et al. (2012). "RBFOX1 regulates both splicing and transcriptional networks in human neuronal development." Hum Mol Genet **21**(19): 4171-4186.

Foshay, K. M. and G. I. Gallicano (2008). "Regulation of Sox2 by STAT3 initiates commitment to the neural precursor cell fate." Stem Cells Dev **17**(2): 269-278.

Frazer, K. A., et al. (2001). "Evolutionarily conserved sequences on human chromosome 21." Genome Res **11**(10): 1651-1659.

Friedman, R. C., et al. (2009). "Most mammalian mRNAs are conserved targets of microRNAs." Genome research **19**(1): 92-105.

Frith, M. C., et al. (2001). "Detection of cis-element clusters in higher eukaryotic DNA." Bioinformatics **17**(10): 878-889.

Fujimoto, S. Y., et al. (2000). "Arabidopsis ethylene-responsive element binding factors act as transcriptional activators or repressors of GCC box-mediated gene expression." Plant Cell **12**(3): 393-404.

Griffiths-Jones, S., et al. (2006). "miRBase: microRNA sequences, targets and gene nomenclature." Nucleic acids research **34**(suppl 1): D140-D144.

Grimson, A., et al. (2007). "MicroRNA targeting specificity in mammals: determinants beyond seed pairing." Molecular cell **27**(1): 91-105.

Grossman, A. R., et al. (2003). "Chlamydomonas reinhardtii at the crossroads of genomics." Eukaryot Cell **2**(6): 1137-1150.

Gupta, M. and J. S. Liu (2005). "De novo cis-regulatory module elicitation for eukaryotic genomes." Proc Natl Acad Sci U S A **102**(20): 7079-7084.

Hafner, M., et al. (2010). "Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP." Cell **141**(1): 129-141.

Hattori, T., et al. (1995). "Regulation of the Osem gene by abscisic acid and the transcriptional activator VP1: analysis of cis-acting promoter elements required for regulation by abscisic acid and VP1." Plant J **7**(6): 913-925.

Helwak, A., et al. (2013). "Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding." Cell **153**(3): 654-665.

Hertz, G. Z. and G. D. Stormo (1999). "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences." Bioinformatics **15**(7-8): 563-577.

Higo, K., et al. (1999). "Plant cis-acting regulatory DNA elements (PLACE) database: 1999." Nucleic Acids Res **27**(1): 297-300.

Hofacker, I. L. (2003). "Vienna RNA secondary structure server." Nucleic acids research **31**(13): 3429-3431.

Hu, J., et al. (2008). "MOPAT: a graph-based method to predict recurrent cis-regulatory modules from known motifs." Nucleic Acids Res **36**(13): 4488-4497.

Hu, M., et al. (2010). "On the detection and refinement of transcription factor binding sites using ChIP-Seq data." Nucleic Acids Res **38**(7): 2154-2167.

Huang, H., et al. (2004). "Determination of local statistical significance of patterns in Markov sequences with application to promoter element identification." J Comput Biol **11**(1): 1-14.

Ideker, T., et al. (2001). "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network." Science **292**(5518): 929-934.

Jayaswal, V., et al. (2011). "Identification of microRNA-mRNA modules using microarray data." BMC Genomics **12**: 138.

Ji, H., et al. (2008). "An integrated software system for analyzing ChIP-chip and ChIP-seq data." Nat Biotechnol **26**(11): 1293-1300.

Jin, V. X., et al. (2009). "W-ChIPMotifs: a web application tool for de novo motif discovery from ChIP-based high-throughput data." Bioinformatics **25**(23): 3191-3193.

John, B., et al. (2004). "Human microRNA targets." PLoS Biol **2**(11): e363.

Johnson, D. S., et al. (2007). "Genome-wide mapping of in vivo protein-DNA interactions." Science **316**(5830): 1497-1502.

Jothi, R., et al. (2008). "Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data." Nucleic Acids Res **36**(16): 5221-5231.

Kent, W. J., et al. (2002). "The human genome browser at UCSC." Genome Res **12**(6): 996-1006.

Kertesz, M., et al. (2007). "The role of site accessibility in microRNA target recognition." Nature genetics **39**(10): 1278-1284.

Khan, A. A., et al. (2009). "Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs." Nat Biotechnol **27**(6): 549-555.

Kim, Y. and J. Kim (2004). Gradient LASSO for feature selection. Proceedings of the twenty-first international conference on Machine learning, ACM.

Kishore, S., et al. (2011). "A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins." Nature methods **8**(7): 559-564.

Kloosterman, W. P., et al. (2004). "Substrate requirements for let-7 function in the developing zebrafish embryo." Nucleic Acids Res **32**(21): 6284-6291.

Kluiver, J., et al. (2005). "BIC and miR‐155 are highly expressed in Hodgkin, primary mediastinal and diffuse large B cell lymphomas." The Journal of pathology **207**(2): 243-249.

Kokaly, R. F. and R. N. Clark (1999). "Spectroscopic determination of leaf biochemistry using band-depth analysis of absorption features and stepwise multiple linear regression." Remote sensing of environment **67**(3): 267-287.

Krek, A., et al. (2005). "Combinatorial microRNA target predictions." Nature genetics **37**(5): 495-500.

Krüger, J. and M. Rehmsmeier (2006). "RNAhybrid: microRNA target prediction easy, fast and flexible." Nucleic acids research **34**(suppl 2): W451-W454.

Kulakovskiy, I. V., et al. (2010). "Deep and wide digging for binding motifs in ChIP-Seq data." Bioinformatics **26**(20): 2622-2623.

Langner, U., et al. (2009). "An energy balance from absorbed photons to new biomass for Chlamydomonas reinhardtii and Chlamydomonas acidophila under neutral and extremely acidic growth conditions." Plant Cell Environ **32**(3): 250-258.

Lawrence, C. E., et al. (1993). "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment." Science **262**(5131): 208-214.

Lescot, M., et al. (2002). "PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences." Nucleic acids research **30**(1): 325-327.

Lewis, B. P., et al. (2005). "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets." Cell **120**(1): 15-20.

Lewis, B. P., et al. (2003). "Prediction of mammalian microRNA targets." Cell **115**(7): 787-798.

Li, J., et al. (2014). "Identifying mRNA sequence elements for target recognition by human Argonaute proteins." Genome research.

Li, L., et al. (2006). "A mixture model-based discriminate analysis for identifying ordered transcription factor binding site pairs in gene promoters directly regulated by estrogen receptor-alpha." Bioinformatics **22**(18): 2210-2216.

Li, W. H., et al. (1985). "A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes." Mol Biol Evol **2**(2): 150-174.

Li, X. and W. H. Wong (2005). "Sampling motifs on phylogenetic trees." Proc Natl Acad Sci U S A **102**(27): 9481-9486.

Li, X., et al. (2005). "Reliable prediction of transcription factor binding sites by phylogenetic verification." Proc Natl Acad Sci U S A **102**(47): 16945-16950.

Li, Y., et al. "Inhibition of starch synthesis results in overproduction of lipids in Chlamydomonas reinhardtii." Biotechnol Bioeng **107**(2): 258-268.

Licatalosi, D. D., et al. (2008). "HITS-CLIP yields genome-wide insights into brain alternative RNA processing." Nature **456**(7221): 464-469.

Liu, X., et al. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pacific symposium on biocomputing, World Scientific.

Liu, X. S., et al. (2002). "An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments." Nat Biotechnol **20**(8): 835-839.

Liu, X. S., et al. (2002). "An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments." Nature biotechnology **20**(8): 835-839.

Liu, Y., et al. (2004). "Eukaryotic regulatory element conservation analysis and identification using comparative genomics." Genome Res **14**(3): 451-458.

Loher, P. and I. Rigoutsos (2012). "Interactive exploration of RNA22 microRNA target predictions." Bioinformatics **28**(24): 3322-3323.

Loots, G. G., et al. (2000). "Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons." Science **288**(5463): 136-140.

Lucker, B. F., et al. (2010). "Direct interactions of intraflagellar transport complex B proteins IFT88, IFT52, and IFT46." J Biol Chem **285**(28): 21508-21518.

Ma, S. and H. J. Bohnert (2007). "Integration of Arabidopsis thaliana stress-related transcript profiles, promoter structures, and cell-specific expression." Genome Biol **8**(4): R49.

Ma, S. and J. Huang (2008). "Penalized feature selection and classification in bioinformatics." Briefings in bioinformatics **9**(5): 392-403.

Mahony, S. and P. V. Benos (2007). "STAMP: a web tool for exploring DNA-binding motif similarities." Nucleic Acids Res **35**(Web Server issue): W253-258.

Mann, H. B. and D. R. Whitney (1947). "On a test of whether one of two random variables is stochastically larger than the other." The annals of mathematical statistics: 50-60.

Maragkakis, M., et al. (2009). "DIANA-microT web server: elucidating microRNA functions through target prediction." Nucleic acids research: gkp292.

Marino-Ramirez, L., et al. (2004). "Statistical analysis of over-represented words in human promoter sequences." Nucleic Acids Res **32**(3): 949-958.

Meinshausen, N. and P. Bühlmann (2010). "Stability selection." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72**(4): 417-473.

Meister, G. (2013). "Argonaute proteins: functional insights and emerging roles." Nature Reviews Genetics **14**(7): 447-459.

Merchant, S. S., et al. (2007). "The Chlamydomonas genome reveals the evolution of key animal and plant functions." Science **318**(5848): 245-250.

Metzler, M., et al. (2004). "High expression of precursor microRNA‐155/BIC RNA in children with Burkitt lymphoma." Genes, Chromosomes and Cancer **39**(2): 167-169.

Miller, R., et al. (2010). "Changes in transcript abundance in Chlamydomonas reinhardtii following nitrogen deprivation predict diversion of metabolism." Plant Physiol **154**(4): 1737-1752.

Mittag, M., et al. (2005). "The circadian clock in Chlamydomonas reinhardtii. What is it for? What is it similar to?" Plant Physiol **137**(2): 399-409.

Muljo, S. A., et al. (2010). "MicroRNA targeting in mammalian genomes: genes and mechanisms." Wiley Interdisciplinary Reviews: Systems Biology and Medicine **2**(2): 148-161.

Nakashima, K. (1999). "Synergistic Signaling in Fetal Brain by STAT3-Smad1 Complex Bridged by p300." Science **284**(5413): 479-482.

Nguyen, M. T., et al. (2009). "Hydrothermal acid pretreatment of Chlamydomonas reinhardtii biomass for ethanol production." J Microbiol Biotechnol **19**(2): 161-166.

Ono, A., et al. (1996). "The rab16B promoter of rice contains two distinct abscisic acid-responsive elements." Plant Physiol **112**(2): 483-491.

Paraskevopoulou, M. D., et al. (2013). "DIANA-microT web server v5. 0: service integration into miRNA functional analysis workflows." Nucleic acids research **41**(W1): W169-W173.

Paraskevopoulou, M. D., et al. (2013). "DIANA-microT web server v5. 0: service integration into miRNA functional analysis workflows." Nucleic acids research: gkt393.

Pardo, M., et al. (2010). "An expanded Oct4 interaction network: implications for stem cell biology, development, and disease." Cell Stem Cell **6**(4): 382-395.

Perez-Rodriguez, P., et al. (2009). "PlnTFDB: updated content and new features of the plant transcription factor database." Nucleic Acids Res **38**(Database issue): D822-827.

Peterson, S. M., et al. (2014). "Common features of microRNA target prediction tools." Front Genet **5**: 23.

Pollard, K. S., et al. (2010). "Detection of nonneutral substitution rates on mammalian phylogenies." Genome research **20**(1): 110-121.

Portales-Casamar, E., et al. (2010). "JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles." Nucleic Acids Res **38**(Database issue): D105-110.

Prochnik, S. E., et al. (2010). "Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri." Science **329**(5988): 223-226.

Que, J., et al. (2007). "Multiple dose-dependent roles for Sox2 in the patterning and differentiation of anterior foregut endoderm." Development **134**(13): 2521-2531.

Ralston, A. and H. S. Wilf (1960). "Mathematical methods for digital computers."

Reczko, M., et al. (2012). "Functional microRNA targets in protein coding sequences." Bioinformatics **28**(6): 771-776.

Reczko, M., et al. (2011). "Accurate microRNA target prediction using detailed binding site accessibility and machine learning on proteomics data." Frontiers in genetics **2**.

Reed, B. D., et al. (2008). "Genome-wide occupancy of SREBP1 and its partners NFY and SP1 reveals novel functional roles and combinatorial regulation of distinct classes of genes." PLoS Genet **4**(7): e1000133.

Robertson, G., et al. (2007). "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing." Nat Methods **4**(8): 651-657.

Rochaix, J. D. (2004). "Genetics of the biogenesis and dynamics of the photosynthetic machinery in eukaryotes." Plant Cell **16**(7): 1650-1660.

Rodriguez, A., et al. (2007). "Requirement of bic/microRNA-155 for normal immune function." Science **316**(5824): 608-611.

Rombauts, S., et al. (1999). "PlantCARE, a plant cis-acting regulatory element database." Nucleic Acids Res **27**(1): 295-296.

Saetrom, P., et al. (2007). "Distance constraints between microRNA target sites dictate efficacy and cooperativity." Nucleic Acids Res **35**(7): 2333-2342.

Saeys, Y., et al. (2007). "A review of feature selection techniques in bioinformatics." Bioinformatics **23**(19): 2507-2517.

Sandelin, A., et al. (2004). "JASPAR: an open-access database for eukaryotic transcription factor binding profiles." Nucleic Acids Res **32**(Database issue): D91-94.

Sassen, S., et al. (2008). "MicroRNA—implications for cancer." Virchows Archiv **452**(1): 1-10.

Schanen, B. C. and X. Li (2011). "Transcriptional regulation of mammalian miRNA genes." Genomics **97**(1): 1-6.

Schmitter, D., et al. (2006). "Effects of Dicer and Argonaute down-regulation on mRNA levels in human HEK293 cells." Nucleic Acids Res **34**(17): 4801-4815.

Segal, E., et al. (2003). "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data." Nat Genet **34**(2): 166-176.

Shannon, P., et al. (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome Res **13**(11): 2498-2504.

Sharov, A. A. and M. S. Ko (2009). "Exhaustive search for over-represented DNA sequence motifs with CisFinder." DNA Res **16**(5): 261-273.

Singh, K. B. (1998). "Transcriptional regulation in plants: the importance of combinatorial control." Plant Physiol **118**(4): 1111-1120.

Sinha, S., et al. (2004). "PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences." BMC Bioinformatics **5**: 170.

Skriver, K., et al. (1991). "cis-acting DNA elements responsive to gibberellin and its antagonist abscisic acid." Proc Natl Acad Sci U S A **88**(16): 7266-7270.

Sokal, R. and C. Michener (1985). "A statistical method for evaluating systematic relationships." University of Kansas Science Bulletin **38**: 1409-1438.

Staden, R. (1989). "Methods for calculating the probabilities of finding patterns in sequences." Comput Appl Biosci **5**(2): 89-96.

Stamatoyannopoulos, J. A. (2012). "What does our genome encode?" Genome Res **22**(9): 1602-1611.

Steffens, N. O., et al. (2005). "AthaMap web tools for database-assisted identification of combinatorial cis-regulatory elements and the display of highly conserved transcription factor binding sites in Arabidopsis thaliana." Nucleic Acids Res **33**(Web Server issue): W397-402.

Storey, J. D. and R. Tibshirani (2003). "Statistical significance for genomewide studies." Proc Natl Acad Sci U S A **100**(16): 9440-9445.

Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.

Stormo, G. D. and G. W. Hartzell, 3rd (1989). "Identifying protein-binding sites from unaligned DNA fragments." Proc Natl Acad Sci U S A **86**(4): 1183-1187.

Sumazin, P., et al. (2011). "An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma." Cell **147**(2): 370-381.

Sun, H., et al. (2012). "Unveiling combinatorial regulation through the combination of ChIP information and in silico cis-regulatory module detection." Nucleic Acids Res **40**(12): e90.

Sun, T. H., et al. "Coordinated regulation of gene expression for carotenoid metabolism in Chlamydomonas reinhardtii." J Integr Plant Biol **52**(10): 868-878.

Svetnik, V., et al. (2003). "Random forest: a classification and regression tool for compound classification and QSAR modeling." Journal of chemical information and computer sciences **43**(6): 1947-1958.

Thai, T.-H., et al. (2007). "Regulation of the germinal center response by microRNA-155." Science **316**(5824): 604-608.

Thomas-Chollier, M., et al. (2012). "RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets." Nucleic Acids Res **40**(4): e31.

Thomson, D. W., et al. (2011). "Experimental strategies for microRNA target identification." Nucleic acids research **39**(16): 6845-6853.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological): 267-288.

Ule, J., et al. (2003). "CLIP identifies Nova-regulated RNA networks in the brain." Science **302**(5648): 1212-1215.

Urzica, E. I., et al. (2012). "Impact of Oxidative Stress on Ascorbate Biosynthesis in Chlamydomonas via Regulation of the VTC2 Gene Encoding a GDP-L-galactose Phosphorylase." J Biol Chem **287**(17): 14234-14245.

Valouev, A., et al. (2008). "Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data." Nat Methods **5**(9): 829-834.

Vandepoele, K., et al. (2006). "Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics." Genome Biol **7**(11): R103.

Vaquerizas, J. M., et al. (2009). "A census of human transcription factors: function, expression and evolution." Nature Reviews Genetics **10**(4): 252-263.

Vejnar, C. E., et al. (2013). "miRmap web: comprehensive microRNA target prediction online." Nucleic acids research **41**(W1): W165-W168.

Vejnar, C. E. and E. M. Zdobnov (2012). "MiRmap: comprehensive prediction of microRNA target repression strength." Nucleic acids research **40**(22): 11673-11683.

Vella, M. C., et al. (2004). "The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR." Genes Dev **18**(2): 132-137.

Vergoulis, T., et al. (2012). "TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support." Nucleic acids research **40**(D1): D222-D229.

Wang, T. and G. D. Stormo (2003). "Combining phylogenetic data with co-regulated genes to identify regulatory motifs." Bioinformatics **19**(18): 2369-2380.

Wang, T. and G. D. Stormo (2005). "Identifying the conserved network of cis-regulatory sites of a eukaryotic genome." Proc Natl Acad Sci U S A **102**(48): 17400-17405.

Wang, T., et al. (2014). "dCLIP: a computational approach for comparative CLIP-seq analyses." Genome Biol **15**(1): R11.

Wang, X. (2014). "Composition of seed sequence is a major determinant of microRNA targeting patterns." Bioinformatics **30**(10): 1377-1383.

Wang, X. and I. M. El Naqa (2008). "Prediction of both conserved and nonconserved microRNA targets in animals." Bioinformatics **24**(3): 325-332.

Wang, X., et al. (2009). "Discovery of cis-elements between sorghum and rice using co-expression and evolutionary conservation." BMC Genomics **10**: 284.

Wang, Y., et al. (2005). "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer." Lancet **365**(9460): 671-679.

Wang, Y., et al. (2011). "Transcriptional regulation of co-expressed microRNA target genes." Genomics **98**(6): 445-452.

Weigelt, B., et al. (2005). "Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer." Cancer Res **65**(20): 9155-9158.

Wemmer, K. A. and W. F. Marshall (2004). "Flagellar motility: all pull together." Curr Biol **14**(23): R992-993.

Wilbanks, E. G. and M. T. Facciotti (2010). "Evaluation of algorithm performance in ChIP-seq peak detection." PloS one **5**(7): e11471.

Wilcoxon, F. (1945). "Individual comparisons by ranking methods." Biometrics Bulletin **1**(6): 80–83.

Wingender, E., et al. (1996). "TRANSFAC: a database on transcription factors and their DNA binding sites." Nucleic Acids Res **24**(1): 238-241.

Witkos, T. M., et al. (2011). "Practical Aspects of microRNA Target Prediction." Curr Mol Med **11**(2): 93-109.

Wu, S., et al. (2010). "Multiple microRNAs modulate p21Cip1/Waf1 expression by directly targeting its 3' untranslated region." Oncogene **29**(15): 2302-2308.

Xu, J., et al. (2011). "MiRNA-miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features." Nucleic Acids Res **39**(3): 825-836.

Yamamoto, Y. Y., et al. (2011). "Prediction of transcriptional regulatory elements for plant hormone responses based on microarray data." BMC Plant Biol **11**: 39.

Yanaihara, N., et al. (2006). "Unique microRNA molecular profiles in lung cancer diagnosis and prognosis." Cancer cell **9**(3): 189-198.

Yang, N., et al. (2008). "MicroRNA epigenetic alterations in human cancer: one step forward in diagnosis and treatment." International journal of cancer **122**(5): 963-968.

Yeo, C. J., et al. (1995). "A prospective randomized trial of pancreaticogastrostomy versus pancreaticojejunostomy after pancreaticoduodenectomy." Annals of surgery **222**(4): 580.

Yousef, M., et al. (2007). "Naïve Bayes for microRNA target predictions—machine learning for microRNA targets." Bioinformatics **23**(22): 2987-2992.

Yu, S.-L., et al. (2008). "MicroRNA signature predicts survival and relapse in lung cancer." Cancer cell **13**(1): 48-57.

Yuh, C. H., et al. (1998). "Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene." Science **279**(5358): 1896-1902.

Zhang, S., et al. (2011). "A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules." Bioinformatics **27**(13): i401-409.

Zhang, W., et al. (2000). "The gut-enriched Kruppel-like factor (Kruppel-like factor 4) mediates the transactivating effect of p53 on the p21WAF1/Cip1 promoter." J Biol Chem **275**(24): 18391-18398.

Zhang, Y., et al. (2008). "Model-based analysis of ChIP-Seq (MACS)." Genome Biol **9**(9): R137.

Zhou, Q. and W. H. Wong (2004). "CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling." Proc Natl Acad Sci U S A **101**(33): 12114-12119.