Computational Biology of Gene Regulation

Lecturer: Dr Alex Stark

Scriber: Fulu Li

1. Introduction

The main topic of this lecture is about how gene expression is regulated. The regulation of gene expression is of utmost importance for both unicellular organisms that need to adapt to environmental changes and nutritional stresses, and for metazoa, in which individual cells respond to extracellular signals such as hormones and growth factors by switching their activities from one state to another or changing their proliferation or differentiation status [8]. Notably, different genes have different expression patterns and the regulation of gene expression is normally accomplished at two major stages: (1) the transcription level via transcription factors/motifs (such as E2F4 transcription factor for cell-cycle regulation); (2) the translation level (also known as post-transcriptional regulation) via microRNAs (such as some microRNA inhibits the E2F1 protein, which regulates cell proliferation [9]). microRNA appears to bind to messenger RNA (mRNA) before it can be translated to proteins that switch genes on and off.

The rest of the scribe notes are organized as follows: we address the basic concepts of motifs and motif discovery methods in Section 2; we discuss the basic concepts of microRNAs and the gene regulation by microRNAs in Section 3. A brief summary is given in Section 4.

2. Motif Discovery

In this section, we address the basic concepts of motif and focus on motif discovery methods, which is the key to obtain insights on the role of regulatory motifs in gene expression.

2.1 Motifs

Many functionally important regions of the genome can be recognized by searching for sequence patterns, or “motifs” and identifying the regulatory motifs bound by transcription factors can provide crucial insight into the mechanisms of transcriptional regulation [11]. One way of representing a motif is by using a consensus sequence of preferred nucleotides. In essence, a motif is simply a short nucleotide sequence of about 6-15 nucleotides in length embedded in a longer DNA sequence. Another way of modeling motif is to use the position weight matrix (PWM), where the motif is represented as a matrix of nucleotide scores indexed by nucleotide and position and it is
often illustrated in a graphical motif logo (see also *Lecture 9* on regulatory motif discovery).

The search for those sites of regulatory motifs is challenging due to the fact that a single regulatory protein will often recognize a variety of similar sequences [11].

### 2.2 Motif Discovery Methods

In general, motifs can be discovered experimentally as well as computationally. Motif discovery is to learn representations of regulatory motifs from sequence data.

Motifs are normally bound by TF. For the experimental determination of TF motifs, we can apply any one of the two principles: (1) hold on the transcriptional factor, i.e., the immobilization of TF; (2) hold on the DNA, i.e., the immobilization of DNA. For the former approach, we have SELEX (systematic evolution of ligands by exponential enrichment) and DIP-chip (DNA-immunoprecipitation with microarray detection to determine binding specificity); for the latter, we have PBM (protein-binding microarrays). The drawback of SELEX method is that it might result in only high fidelity binders.

For the computational discovery of motifs, we have enumerative approaches and alignment methods (such as EM -- expectation maximization, Gibbs sampling, etc., see also *Lecture 9* for details about these methods). Enumerative approaches typically involve exhaustive enumeration of words up to a certain maximum length in a data set, which is better suited to consensus sequence motif models [11]. On the other hand, alignment methods often involve development of a probabilistic model of the observed sequence data and optimization to find motifs common to all input sequences [11]. The drawback for those computational approaches is that it requires sensible grouping of genes/promoters.

In general, there are two major types of approaches for the discovery of motifs: overrepresentation method and conservation method. The enumeration approaches and alignment methods are all overrepresentation-based schemes. Conservation-based methods assume that known motifs are frequently conserved [4]. However, by looking only for sequences that have remained roughly the same, one would miss a large number of functional elements. For instance, protein-coding genes can undergo extensive changes and yet retain their critical functions [7].

### 2.3. A Systematic Process

The process to systematically test candidate motif patterns is as follows: (1) enumerate all potential motifs; (2) evaluate motif conservation score (MCS); (3) cluster similar motifs; (4) output final motifs. One caveat to make MCS work across the whole genome is to adjust the confidence intervals to more conservative lower ends, which is measured by motif-excess-score (MEC) [4]. Transcriptional factor motifs often have preferential binding sites. Regulatory motifs usually have exclusion, clustering and positional
constraints. The functional evidence for some promoter motifs include the demonstration of positional bias, tissue-specific expression in target genes as well as the matching of known motifs. Notably, this systematic process is also in the category of overrepresentation-based approach.

2.4. Challenges with Large Phylogeny

Large phylogeny poses great challenges for motif discovery due to some sequencing/alignment artifacts (such as low coverage sequencing and mis-alignments, etc.) as well as evolutionary variations (such as the movement/mutation of individual binding sites, selected conversation only in subset of species, etc.). To tackle these challenges, a branch-length cutoff approach is employed without the requirement of perfect conservation. Moreover, window-based search is applied without the requirement of exact alignment. As such, the branch-length-score (BLS) measure is performed to assess conservation to account for errors in sequencing, assembly, and alignment. Notably, the conservation of transcriptional factor motifs and the conservation of miRNA binding motifs select different regions with functional instances.

2.5. Motif Discovery in Practice

A comprehensive review on motif discovery in practice is presented in [11], which is mainly about over-representation approaches.

- Motif discovery typically begins with a group of putatively co-regulated genes, which are often obtained via clustering.
- A motif with very low information content is difficult to distinguish from the background sequence. Basic statistical considerations relating to motif frequency and overrepresentation in the dataset also affect performance.
- Using multiple motif discovery programs often improves performance.
- It is also critical to use a consistent scoring metric that allows motifs to be compared and ranked regardless of their sources.
- Clustering motifs makes the analysis easier.
- Empirical significance testing and cross-validation reduce the risk of overfitting.
- Phylogenetic conservation information improves motif discovery performance.

3. Gene Regulation by miRNAs

In this section, we focus on the gene regulation by miRNAs, which were discovered only a decade ago [13].

3.1. microRNAs

microRNAs (miRNAs) are single-stranded RNA molecules of about 21-23 nucleotides in length, which regulate gene expression. miRNAs are encoded by genes that are transcribed from DNA but not translated into protein (non-coding RNA). In essence, microRNAs are short segments of RNA that regulate gene expression by binding to target
messenger RNAs (mRNAs) and by controlling protein production or causing RNA cleavage [3].

3.2. microRNA as a major regulator

Since their discovery a decade ago [13], microRNAs have emerged as major regulators of gene expression. As mentioned before, a miRNA binds to a specific target sequence within a much longer messenger RNA (mRNA), inhibiting its translation and thus controlling expression of the corresponding gene even after the DNA itself has been read. Within the human genome, there are about 533 genes that code for miRNAs [14]. The important role of miRNAs in animals is highlighted by recent estimates that 20%-30% of all genes are miRNA targets [2].

An individual miRNA may regulate multiple genes and the genes regulated by a single miRNA may be functionally related, such as components of the protein degradation system or specific signal transduction pathways [10]. It is reported that an individual miRNA may regulate as many as hundreds of genes [15] and most known targets of genetically identified miRNAs contain multiple sites [2]. However, 95% of genes with conserved target sites have just one site for one miRNA [2].

3.3 Identifying miRNA Targets

To efficiently identify miRNA targets, we need to better understand the characteristics of miRNA binding sites, seed pairing, possible modes of miRNA action, and molecular characteristics of targets and anti-targets, etc.

Originally, genetically involved miRNA targets show that miRNA-target duplexes contain: (a) limited sequence complementarity; (b) gaps and mismatches at variable positions; (c) variable number of G:U base-pairs. All those revelations indicate that target identification in animal genomes is not possible by standard sequence comparisons methods [3]. It is a complex issue involved with both the problem of sensitivity and the problem of specificity.

The algorithms to evaluate whether a potential target site is likely to be actually regulated by miRNA are based on several factors [3,5,6,10]: (a) the importance of 5’ pairing has to be considered (we refer interested readers to [15] for details); (b) the target site must have some degree of sequence complementary to one or more of the known miRNAs; (c) the strength with which the predicted target and its miRNA bind together (e.g., the binding strength), which can be calculated from the sequence and other structural factors, must be higher than some threshold; (d) evolutionary conservation and evolutionary signature — the presence of the miRNA-target pair in different organisms and different genomes of closely-related species — is factored in, because the likelihood that the miRNA and target actually pair in vivo is greater if the pair is found in multiple types of organisms and multiple genomes of closely-related species [3,5,6,10]; In general, the basic premise of comparative genomics is that if something is conserved over millions of years in several closely-related species, it's likely to do something useful [7].
3.4. Identifying miRNAs

There is enormous computational challenge to correctly identify miRNA. For example, in fly genome there are about 760,355 random hairpins, among which a handful of them are mapped to 60 or so miRNAs. If the false positive rate is only 0.5%, there will be 3800 spurious hairpins. Therefore, a specificity rate has to be greater than 99.99% in order to make reasonable predictions.

Although structural features of miRNAs provide some discriminatory power, structure alone can not reliable predict miRNAs. We have to resort to evolutionary signatures (such as miRNA hairpin characteristics, region specificity, seed conservation and avoidance, miRNA annotation, etc.) as well as experimental validation for reliable prediction and verification of miRNAs. We need to emphasize that the evolutionary signature of miRNAs is often described as the high arm and low loop conservation model, which is also known as the “camel” model [4].

3.5 The Roles of miRNAs in Gene Expression

First of all, there are many roles of miRNAs in gene expression yet to be discovered due to the fact that gene expression was thought to be regulated mainly at the transcriptional level for several decades and miRNAs were discovered only a decade ago.

Analysis of miRNA target sites reveals that miRNAs control transcriptional noise (such as some mutually exclusive expression patterns) [2,15].

Many genetically identified examples show that miRNAs function as gene switches via the binding to mRNA (messenger RNA) before it is translated into protein. For example, $\text{lin-4}$ and $\text{let-7}$ miRNAs control developmental timing [13, 17] and $\text{lsy-6}$ miRNA regulates left–right asymmetry in the nervous system [18]. It is also reported that mouse $\text{miR-181}$ is preferentially expressed in bone marrow and was shown to be involved in hematopoietic differentiation [19].

Some recent studies discovered emerging roles of miRNA to control the threshold levels in feed-back loops. For example, miRNA-9a ensures the precise specification of sensory organ precursors in Drosophila [16].

3.6. Implications of Recent Findings

In this section, we briefly discuss the significant implications of recent findings in gene regulation, in particular on the emerging roles of regulatory motifs and miRNAs via the comparative study of a number of recently-sequenced genomes.

Recent analysis [1,4,5,7] of a dozen newly-sequenced fly genomes showed that some microRNA genes can produce many functional products from single regions, encoded in
overlapping ways. This is an entirely new mechanism and the researchers have since found evidence of this mechanism in the human genome as well.

4. Summary

Gene expression is controlled by specific interactions between regulatory proteins, transcription factors and short sequences in the regulatory regions of genes to which they bind. Control regions are modular and the regulatory output of a sequence depends on the specific combination of its elements as well as, partially, on the order and on the orientation in which they occur [12].

Regulatory motifs and miRNAs both play important regulatory roles in gene expression. As mentioned before, many functionally important regions of the genome can be recognized by searching for sequence patterns, or motifs. It is reported that miRNA genes are one of the most abundant classes of regulatory genes in animals, estimated between 0.5% and 1% of the predicted genes in worms, flies, and humans, raising the prospect that they could have many more regulatory functions than those discovered to date [6]. Indeed, recent studies in [1,4,5,7] have found thousands of previously unidentified functional elements, including more than a hundred miRNA genes.

In summary, gene regulatory processes (e.g. motif regulation, mRNA processing, export, surveillance, silencing and turnover) are interlinked by the use of common factors and constitute a complex regulatory network that contributes to cell-type and organism specific gene expression patterns [1,4,8]. The tremendous research efforts and discoveries in this frontier will pave the way for a better understanding of human genome and a better understanding of human regulatory mechanisms at large.

Reference


