Statistical models of non-coding RNA-mediated gene regulation

Annalisa Marsico
Max Planck Institute for Molecular Genetics, Free University of Berlin
marsico@molgen.mpg.de

1 Abstract

In our group, called RNA Bioinformatics, we are interested in investigating regulation and function of non-coding RNA transcripts by means of *in silico* methods. In particular, we are interested in those non-coding RNAs which act as regulators of gene expression, and their interplay with Transcription Factors (TFs), epigenetic marks and RNA Binding Proteins (RBPs). High-throughput experiments provide a rich source of information: the integration and interpretation of different genomic data requires the development of adequate statistical models and algorithms to uncover the putative role of non-coding transcripts in regulatory network. Regularized or sparse regression models are the main methods we employ in order to derive mechanistic hypothesis about non-coding RNA function, which can be later tested in the wet lab. We further apply unsupervised methods, such as k-means clustering or spectral clustering to characterize the properties of new non-coding RNA sub-classes by integrating several sources of high-throughput genomic data.

2 Introduction

In the cell, genomic DNA is transcribed into various types of RNA, but not all RNAs are translated into proteins. Over the past few years it has been observed, thanks to high-throughput sequencing methods, that a big portion of the human genome is transcribed in a tissue- and time-specic manner [ea13c]. Most the detected transcripts in mammals and other complex organisms are non-coding RNAs (ncRNAs), RNAs that do not encode for proteins [Mat06]. Although the functional consequences of different ncRNA classes are not yet fully understood, this does not mean that they do not contain information nor have functions.

Among the different classes of non-coding transcripts, microRNAs (miRNAs), small RNAs of 18 to 24 nucleotides in length, that post-transcriptionally regulated gene expression, are the most widely studied, but other classes of small non-coding RNAs have been characterized, such as snoRNAs (many of which still remain to be identified), snRNAs and piRNAs [ea12b]. At post-transcriptional level, about half of the human genes are regulated by microRNAs (miRNAs), which can bind to the 3-untranslated regions (3' UTRs) or coding regions of target genes, leading to the degradation of target mRNAs or translational repression [ea14a]. MiRNAs are associated with an array of biological processes, such as embryonic development and stem cell functions in mammals [ea09], and a crucial role of miRNAs in gene regulatory networks has been recognized in the last decade in the context of cancer development, as well as immune system response [ea06]. Given the growing importance of miRNA function in contributing to the control of gene expression, most of the research in the past decade has been focusing on miRNA-gene target prediction and gene regulatory networks have been expanded to include the involvement of miRNAs.

Despite great progress in understanding the biological role of miRNAs, our understanding of how miRNAs are regulated and processed is still developing. High-throughput sequencing data have provided a robust platform for transcriptome-level, as well as gene-promoter analyses. Some recent *in silico* predictive models for miRNA promoter recognition enable the challenging task of locating the Transcription Start Sites (TSSs) of transient miRNA primary transcripts, thereby allowing the prediction of their

regulatory elements and Transcription Factor Binding Sites (TFBSs) [ea11a, ea13a, ea14b].

Besides small non-coding RNAs, advances in high-throughput sequencing, combined with genome-wide mapping of chromatin modification signatures and bioinformatics pipelines, have resulted in the identification of tens of thousands of longer non-coding transcripts (including intergenic and overlapping sense and antisense transcripts), whose functional significance is still controversial [ea15]. Although the existence of such non-coding RNAs has been validated in multiple experimental systems, and several long intergenic non-coding RNAs (lincRNAs) have been described in processes of gene silencing [ea10b], imprinting [ea10c], and lately in gene activation [ea10d, ea13b], a global spectrum of all possible lincRNA-related functions, as well as automated methods for lincRNA function prediction are missing. This is mainly due to the fact that the vast majority of lincRNAs shows little evidence of evolutionary conservation at sequence level [ea11b]. The development of systematic statistical methods to find significant associations or co-expression between protein-coding genes and lincRNAs, as well as the inspection of evolutionary signatures based on structure rather than sequence motifs is needed to help inferring many still-unknown lincRNA functions.

3 Research concept

In our group we are interested in the mechanisms that govern the regulation of non-coding RNAs, as well as functional analysis of different classes of non-coding RNAs, with focus on miRNAs and lincRNAs. To these goals, we develop statistical models and use existing machine learning approaches to answer questions like: How can miRNA promoters be detected genome-wide and distinguished from transcriptional noise? What are the genomic features that control miRNA processing? What are the genes regulated by a certain long non-coding RNA? How can we automatically classify long non-coding RNAs into functional classes based on sequence / structure features? Part of the group focuses on statistical modeling of miRNA regulatory elements from next-generation sequencing data, such as expression data, epigenetics marks and genetic variants that control miRNA expression. The other part of the group focuses on the characterization of lincRNA function using methods, such as sparse regression, network analysis and data integration.

3.1 Statistical modeling of miRNA Biogenesis

MiRNAs are regulated at different levels during their biogenesis pathway [Mat06]. Understanding which TFs regulate a certain miRNA at a certain time in a certain tissue requires the knowledge of the location of the core promoter of the miRNA primary transcript. In my previous work, I developed a semi-supervised machine learning method for miRNA promoter recognition called PROmiRNA [ea13a]. The PROmiRNA mixture model assumes that the read count distribution observed from deepCAGE data is represented by a mixture of putative promoters versus background noise. In order to not underestimate the number of true promoters (due to lowly expressed transcripts), sequence features, such as CpG content, conservation and TATA box affinity are introduced into the model through an informative prior. During training, known miRNA promoters are included as exact examples and the output of the classifier is a posterior probability for a certain regions to be a real promoter. The application of PROmiRNA to the human genome allowed us for the first time to study the characteristics of regulatory elements of different miRNA promoter classes. We are currently working on a more scalable version of the PROmiRNA software, as well as a web-server for allowing the exploration of miRNA promoters across different tissues. By exploiting the functionalities of PROmiRNA we are currently able to explore other aspects of miRNA regulation. Ongoing projects in the lab include 1) the development of a regression model (elastic net) for predicting miRNA expression based on chromatin signatures around at the predicted promoter regions and 2) the prediction of causal genetic variants (eQTL SNPs) which alter miRNA expression in promoter regions by using different regulatory elements as covariates.

Preliminary results indicate a crucial role of DNA methylation in shaping miRNA expression and provide a list of causal genetic variants localized in proximity of tissue-specific miRNA promoters.

Global mature miRNA expression is not only regulated at transcriptional level, but several post-transcriptional steps influence the final miRNA expression level. In our previous work, together with our experimental partners from the group of Dr. Ulf Orom, we have defined a quantitative measure of miRNA processing from RNA-Seq data and built a classification model to discriminate efficient from non-efficient processing based on sequence features, i.e. specific and degenerate k-mers [ea14c]. Prompted by the results of this study we are currently investigating the relationship between miRNA processing and epigenetic signatures of miRNA genes.

3.2 Function prediction of long non-coding RNAs

Recent studies have reported enhancer functions of long non-coding RNAs [ea10d], pointing to active transcription of previously identified enhancer regions. In order to identify specific signatures in this new class of enhancer lincRNAs, in one of our current projects we have collected transcriptome data and epigenetics marks in MCF7 cells and identified, by means of k-means clustering, about 400 lincRNAs with putative active enhancer function. We could also associate this cluster of lincRNAs to high hypomethylation specificity among cell lines and to high co-variation in the expression of their nearby genes, supporting further their role as putative enhancers activated by a methylation-dependent mechanism. Motivated by the fact that if the expression of a gene and a long non-coding RNA co-vary among several tissues, then a direct or indirect association can be inferred between the two, we try to infer putative top ranking associations between genes and long non-coding RNAs across different tissues, Given that the number of variables (all annotated genes and lincRNAs) is much higher than the number of samples (different tissues with available expression data) we use sparse regression techniques, such as Orthogonal Matching Pursuit to prioritize significant interactions.

Long ncRNAs have been shown to physically connect the genomic regions of regulated genes with their own genomic locus, thereby mediating gene activation or enhancer function through direct chromatin interactions [ea10d]. Such data are useful to detect direct interactions, therefore we are currently integrating freely available chromatin-conformation data, such as ChIA-PET data [ea10a], with chromatin states to build the physical interaction network involving genes, long non-coding RNAs and other putative regulatory elements in a specific tissue. Such retrieved interactions can be converted to a weighted adjacency matrix, which we then analyze by means of Spectral Clustering in order to identify potentially important regulatory modules which involve lincRNAs.

Long non-coding RNAs do not act alone to perform their activating or repressing function but often associate with RNA-binding proteins or chromatin remodeling complexes that guide them to their sites of action. Identifying which proteins bind to a specific long non-coding RNA can help shedding light on its function. Interactions of long non-coding RNAs with RNA-binding proteins can be detected via technologies such as CLIP-seq [ea12a]. The technology is really new and so far few methods have been developed to reliably identify binding sites above noise and in the presence of appropriate control, in particular for iCLIP data [ea12a]. Although this project just started, our idea is to model the read count distribution for a certain experiment and the control simultaneously by means of a factorial Hidden Markov Model, taking into account special features of iCLIP experiments (e.g. truncation rates, sequence bias) as additional covariates.

4 Outlook

In summary, our group is working towards a global understanding of how non-coding RNAs, such as miRNAs and long non-coding RNAs, participate in gene regulatory networks. We employ several

machine learning methods, such as semi-supervised or supervised classification models to characterize promoters and regulatory features of miRNAs. Together with the group of Bernd Schmeck at the Uniklinukum Marburg (SFB TR84), we will apply our model in the context of infectious diseases, to elucidate the regulatory mechanisms of miRNAs induced in the host cells by a specific infectious process. Our analysis be will extended to include the possible role of lincRNAs in shaping the regulatory network activated by the host in response to the pathogenic infection, with the hope to discover new functions for long non-coding RNAs. In order to unravel the mechanisms of long non-coding RNA function we would like to discover structural motifs, as well as common RNA-binding protein sites among long non-coding RNAs with similar expression/activation patterns. If we can find signatures or structure/sequence motifs among 'related' lncRNAs, these patterns could hint to the lncRNA function and would represent a first step towards a systemathic functional classification of long non-coding RNAs.

References

[ea06] A. Esquela-Kerscher et al. Oncomirs - microRNAs with a role in cancer. *Nat Rev Cancer*, 6:259–269, 2006.

[ea09] D.P. Bartel et al. MicroRNAs: target recognition and regulatory functions. *Cell*, 136:215–233, 2009.

[ea10a] G. Li et al. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genom Biol*, 11(1):R22, 2010.

[ea10b] M. Huarte et al. A large intergenic non-coding RNA induced by p53 mediates global gene expression in p53 response. *Cell*, 142(3):409–419, 2010.

[ea10c] M. Huarte et al. Long noncoding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, 464(7291):1071–1076, 2010.

[ea10d] U.A. Orom et al. Long noncoding RNAs with enhancer-like function in human cells. *Cell*, 143(1):46–58, 2010.

[ea11a] C. Chien et al. Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data. *Nucl Acids Res*, 39(21):9345–56, 2011.

[ea11b] I. Ulitsky et al. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, 147(7):1537–50, 2011.

[ea12a] J. Koenig et al. Protein-RNA interactions: new genomic technologies and perspectives. *Nat Rev Genet*, 13(2):77–83, 2012.

[ea12b] M.S. Kowalczyk et al. Molecular biology: RNA discrimination. *Nature*, 482:310–311, 2012.

[ea13a] A. Marsico et al. PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs. *Genome Biol*, 14:R84, 2013.

[ea13b] E.G. Berghoff et al. Evf2 (Dlx6as) lncRNA regulates ultraconserved enhancer methylation and the differential transcriptional control of adjacent genes. *Development*, 140:4407–4416, 2013.

[ea13c] I. Ulitsky et al. lincRNAs: Genomics, Evolution, and Mechanisms. *Cell*, 154(1):26–46, 2013.

[ea14a] B.N. Davis et al. Regulation of MicroRNA Biogenesis: A miRiad of mechanisms. *Commun Signal*, 10(7):7–18, 2014.

[ea14b] G. Georgakilas et al. microTSS: accurate microRNA transcription start site identification reveals a significant number of divergent pri-miRNAs. *Nat Commun*, page doi:10.1038/ncomms6700, 2014.

[ea14c] T. Conrad et al. Microprocessor activity controls differential miRNA biogenesis In Vivo. *Cell Rep*, 9:542–554, 2014.

[ea15] J.S. Mattick et al. Discovery and annotation of long noncoding RNAs. *Nat Struct Mol Biol*, 22(1):5–7, 2015.

[Mat06] J.S. Mattick. Non-coding RNA. *Hum Mol Genet*, 15(Supp1):R17–R29, 2006.