Report from Dagstuhl Seminar 19342

Advances and Challenges in Protein-RNA Recognition, Regulation and Prediction

Edited by

Rolf Backofen[1], Yael Mandel-Gutfreund[2], Uwe Ohler[3], and Gabriele Varani[4]

1 **Universität Freiburg, DE,** backofen@informatik.uni-freiburg.de
2 **Technion – Haifa, IL,** yaelmg@technion.ac.il
3 **Max-Delbrück-Centrum – Berlin, DE,** uwe.ohler@mdc-berlin.de
4 **University of Washington – Seattle, US,** varani@chem.washington.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 19342 "Advances and Challenges in Protein-RNA Recognition, Regulation and Prediction".

Seminar August 18–23, 2019 – http://www.dagstuhl.de/19342
2012 ACM Subject Classification Mathematics of computing → Probability and statistics, Theory of computation → Design and analysis of algorithms, Theory of computation → Theory and algorithms for application domains, Applied computing → Life and medical sciences, Applied computing → Chemistry, Applied computing → Mathematics and statistics
Keywords and phrases Machine learning, algorithms, genomics analysis, gene expression networks, big data analysis, quantitative prediction, proteins, RNA, CLIP-Seq
Digital Object Identifier 10.4230/DagRep.9.8.49
Edited in cooperation with Florian Heyl, Michael Uhl

1 Executive Summary

Rolf Backofen
Yael Mandel-Gutfreund
Uwe Ohler
Gabriele Varani

DNA is often described as the blueprint of life, since it encodes all the information necessary for an organism to develop and maintain its biological functions. Single blueprints for specific functions are stored inside DNA regions called genes. The primary product produced (also termed expressed) from genes is RNA, which can either become biologically active itself (non-coding RNA or ncRNA) or is further translated into proteins (messenger RNA or mRNA), which then executes the gene functions. Given the astonishing complexity of biological functions, it is not surprising that the regulation of gene expression itself is a highly complex matter. Proteins, RNA, and DNA all can interact with each other, forming regulatory networks in order to control the expression of genes. To elucidate these networks, experimental scientists rely more and more on high-throughput methods, producing vast amounts of raw data. Computational methods to analyze these huge datasets are therefore of highest demand. The main focus of this seminar lies on RNA-protein and RNA-RNA

Advances and Challenges in Protein-RNA Recognition, Regulation and Prediction, *Dagstuhl Reports*, Vol. 9, Issue 8, pp. 49–69
Editors: Rolf Backofen, Yael Mandel-Gutfreund, Uwe Ohler, and Gabriele Varani
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

interactions. In particular, transcriptome-wide binding patterns of RNA-binding proteins (RBPs), their computational predictability, and the biological effects of binding are discussed. Moreover, the seminar dealt with topics like combinatorial RBP binding prediction, RBP binding kinetics, RNA-RNA interaction prediction, subcellular RNA imaging, and RBP binding site classification. Regarding the computational methodology, several newly developed deep learning methods are presented, e.g. for RBP binding site prediction. Taken together, the aim of the seminar is to bring experimental and computational scientists together for the aforementioned topics and to engage them in fruitful discussions in order to:

- present the current experimental and computational methodologies,
- understand their implications, strengths, and limitations from first-hand experience,
- and spark ideas for developing new computational and experimental methods and improving on existing ones.

2 Table of Contents

3 Introduction

3.1 Seminar Format

The seminar "Advances and Challenges in Protein-RNA Recognition, Regulation and Prediction" emerged from the organizer's combined experience that new experimental and computational methods to understand RNA-based gene regulation are positively influenced and more prolific if both sides exchange their ideas, findings, and hypotheses in order to answer critical biological, biomedical, and bioinformatical questions. Experimental and computational biology are two big fields that require different expertise, which are even further divided into sub-fields such as structural and chemical biology. To foster the research of highly interdisciplinary fields, such as the study of protein-RNA interactions, it is imperative that all sides talk and discuss to understand in-depth the underlying problems and goals to find concrete paths. The biggest opportunities are collaborative analyses combining newly and different computational and experimental methods, but events and venues that facilitate an open platform to form such collaborations and opportunities are missing.

The Dagstuhl seminar brought experimental and computation biologists together, allowing them to present and discuss their findings and newly developed powerful computational and experimental methods to investigate RNA-protein interactions across the genome and transcriptome of different organisms, cell tissues and cells. Each day and each session consisted of a different higher-order topic that combined several presentations of computational and experimental groups to intertwine both sides and flourish vivid discussions. Each session was intermittent by a discussion round to talk about open questions, ideas, and problems for the current topic. These discussions also catalyzed the birth of new and enhanced technologies to improve the analysis of protein-RNA interactions. The seminar was attended by many leading scientists in their field of expertise from all around the world to find solutions and ideas of ongoing problems.

3.2 Studying protein-RNA interactions

In order to understand the complexity of post-transcriptional regulation by RBPs, it is essential to have experimental methods for detecting RBP binding sites on RNAs with high resolution. In this context, CLIP-seq (cross-linking and immunoprecipitation followed by next generation sequencing) together with its popular variants PAR-CLIP, iCLIP, and eCLIP has become the state-of-the-art procedure for determining transcriptome-wide binding sites of RBPs with single-nucleotide resolution.

The seminar saw both the presentation of a new iCLIP variant as well several prediction methods that utilize CLIP-seq data to train models for predicting new binding sites for RBPs of interest. The protocol called in vitro iCLIP makes it possible to identify RBP binding sites in vitro for selected RNAs, which when also taking into account in vivo iCLIP data allows for estimating the effects of trans-acting RBPs on the binding patterns of the studied RBP. The shown prediction methods offer various deep learning approaches to RBP binding site prediction, using, for example, convolutional neural networks (CNNs) or graph convolutional neural networks (GCNs) to learn features from RBP binding sites determined by RNAcompete or CLIP-seq methods.

3.3 Functional analysis of RBPs

Several studies on newly described functions of RBPs or the discovery of RBPs as contributors to certain cellular functions have been presented. Among these were, for example, reports on the ZFP36 family of RBPs with potential functions in steroidogenesis and hypertension, roles of the DAZL RBP in gametogenesis, or a translational repression complex made up of PUM2 and 4E-T active in neuronal specification. Computational approaches have also been shown, for example, the prediction of binding preferences of an RBP based on its protein sequence, which could be used to roughly classify the RBP based on RBPs with known functional roles that share similar binding motifs.

3.4 Study of non-coding RNAs

It is currently estimated that the number of non-coding RNA genes is higher than the amount of protein-coding genes in the human genome. On the other hand, the vast majority of these non-coding RNAs have no assigned functions yet, urging the need for functional studies. In particular, long non-coding RNAs (lncRNAs) have drawn much attention in recent years due to their diverse cellular roles, such as RNA-DNA triple helix formation, or in general DNA interactions to control gene expression. In this context, a work on lncRNAs that bind to transcription factors (TFs) has been presented, linking lncRNA NORAD with TF STAT3 in human stem cells. A second work presented a computational prediction method based on RNA structure alignment and structure-based clustering to identify novel ncRNA-RNA or ncRNA-protein interactions in the human genome. In a third presentation, new functional roles for small nucleolar RNAs (snoRNAs) have been discussed, and a machine learning approach was shown to predict known and new RNAs targeted by snoRNAs.

3.5 Improvement of CLIP-seq Data

The genome-wide approaches reported in Session 1 (CLIP-seq) determine *in vivo* interactions. An interesting discussion evolved about the noise of CLIP-seq experiments. New approaches where discussed and presented to reduce the noise of CLIP-seq experiments, including a better normalization, improved CLIP-seq protocols, and new methods to improve the peak calling quality, such as, improved peak calling algorithms and new methods to reduce the number of false positives and false negatives.

3.6 Provision of Tools and Data for and of protein-RNA experiments

The last vivid discussion session picked up at aforementioned topics about a better contribution and provision of bioinformatical tools for the analysis of protein-RNA experiments, and obtained results of completed analysis of protein-RNA data. A big part of this discussion was Galaxy as an interactive platform for bioinformatical tools. The participants uttered their wishes and ideas about new tools and training materials for Galaxy, such as, IntaRNA, or omniCLIP. Furthermore, the participants discussed about an integrative browser for CLIP-seq results for a quick visualization of the binding sites of different RBPs to improve the investigation of common sequence and structural motifs.

3.7 Conclusions

The third Dagstuhl seminar achieved it's goal to bring computational and experimental people together to exchange and discuss the advances and challenges in protein-RNA recognition, regulation and prediction, which is a very interdisciplinary field and thus challenging from both the experimental and computational standpoint. There are rare occasions and venues that facilitate a platform for open discussions and presentations between experimental and computational people to foster the research of protein-RNA interactions. Consequently, the third Dagstuhl seminar was another successful bridge, which was well appreciated by all participants and gave birth to new and exciting collaborations, ideas, and relationships that would not have been formed without this meeting.

4 Overview of Talks

4.1 How to make Sense out of CLIP-seq data

Rolf Backofen (Universität Freiburg, DE)

It is becoming increasingly clear that a RNA-binding proteins are key elements in regulating the cell's transcriptome. CLIP-seq is one of the major tools to determine binding sites but sufferes from high false negative rate due its expression dependency. This crictical hinders the use of public CLIP-data. We will show in several examples how use of raw public CLIP-seq data can lead to false biological reasoning and how advanced advanced machine learning approach can overome this problem. I will also introduce some recent application of approach that allows us to determine mechanism underlying post-transcriptional regulation. I will further dicuss our results from our new Nature paper, showing that the human RNA helicase DHX9 predomently binds to IRAlu elements and such suppresses the negative effect of Alu inflation in transcipts.

4.2 The kinetic landscape of Dazl-RNA binding in cells

Eckhard Jankowsky

The kinetics by which RNA binding proteins (RBPs) interact with their cellular RNA sites are thought to be critical for the biological function of RBPs, but it has not been possible to measure these kinetic parameters in cells. Here, we describe a new approach to determine kinetic parameters of protein binding to individual RNA sites in cells and show how kinetic data quantitatively link RNA binding patterns to biological RBP function.

We combine time-resolved, multi-photonic RNA-protein crosslinking with Immunoprecipitation, Next Generation Sequencing, and large scale kinetic modeling to determine rate constants for association, dissociation, crosslinking and fractional occupancy for thousands of individual binding sites of the RBP Dazl in mouse GC1 cells. Association and dissociation rate constants for Dazl vary by several orders of magnitude among different binding sites.

Dazl resides at individual binding sites at most for few seconds or less, indicating exceptionally high dynamics of Dazl-RNA binding. We further find that the presence of only few Dazl proteins on a given RNA per time interval ultimately determines the impact of Dazl on translation and decay of a given RNA. Dazl presence on an RNA is controlled in a complex fashion through the binding kinetics at individual binding sites, the collective kinetics of Dazl clusters and the combination of these clusters on a given RNA. The data explain how similar Dazl effects on translation can be accomplished by distinct Dazl-RNA binding patterns. Collectively, our results show that and how previously inaccessible, kinetic parameters for RNA-protein interactions in cells allow the development of detailed mechanistic models for cellular RNA-protein interactions.

4.3 RNA structure as mediator of cooperative/antagonistic RBP interaction

Jörg Fallmann (Universität Leipzig, DE)

Joint work of Jörg Fallmann, Julian König, Katharina Zarnack

RNA is a molecule known for its ability to form stable secondary structures on an inter- and intramolecular level. Such structures can influence the binding site of an RBP. The probability for the formation of basepairs on intramolecular level can be expressed as the accessibility of a stretch of RNA. This accessibility directly influences the availability of binding sites for an interacting molecule. If one RBP binds its target site, this can also lead to changes in the conformation of the RNA molecule, a feature we can model in silico via constraint folding. This approach is used to predict positive and negative effects of RBP interactions on the accessibility of the binding sites for other RBPs or RNAs. From changes in accessibility we can via the calculation of pseudo energy contributions derive a measure for the affinity of the corresponding interaction partner. With this we want to model and infer potential antagonistic or cooperative behavior of RBP pairs or other pairs of interaction partners.

4.4 Associating non-coding RNAs to proteins: from RNA structure to literature mining

Jan Gorodkin (University of Copenhagen, DK)

In the first strategy, we performed a genome-wide prediction of Conserved RNA Structures (CRSs) using the syntenic regions that have been subjected to RNA structural alignment in the mammalian genomes. The downstream integrative analysis of the 774 K predicted CRSs showed that the resulting CRSs are enriched to overlap protein binding sites from CLiP data. Using RNA structure based clustering approach we cluster the CRSs from UTRs and identify thousands of putative CRS clusters that contained at least two CRSs located at different genomic positions. Upon filtering these clusters for consistent overlap to the same protein, we identify a few hundred clusters associating with proteins. In the second strategy, we employ

search for non-coding RNA (ncRNA) – RNA interactions and ncRNA –protein interaction in a similar fashion as done in the STRING database for protein-protein interactions. We use an integrative scoring scheme to obtain confidence scores from four channels: curated examples, experimental data, interaction predictions and automatic literature mining. To evaluate the method we show for the largest class of ncRNAs, microRNAs, that the combined scoring scheme outperform that of individual microRNA target predictors. The obtained interactions link directly into STRINGs payload mechanisms and hence allowing uses to fuse the ncRNA interactions with protein networks.

4.5 Evaluation and Classification of Peak Profiles for Protein-RNA Binding Predictions

Florian Heyl (Universität Freiburg, DE) and Rolf Backofen (Universität Freiburg, DE)

Main reference Florian Heyl, Rolf Backofen: "StoatyDive: Evaluation and Classification of Peak Profiles for Sequencing Data", bioRxiv, Cold Spring Harbor Laboratory, 2019.
URL https://doi.org/10.1101/799114

The prediction of binding sites (peak calling) is a common task in the data analysis of methods such as crosslinking immunoprecipitation in combination with high-throughput sequencing (CLIP-Seq). The peaks are often further analyzed to predict sequence motifs or structure patterns. However, the obtained peak set can vary in their profile shapes because of the used peakcaller method, different binding domains of the protein, protocol biases, or other factors. Thus, a prior step is missing to evaluates and classifies the predicted peaks based on their shapes. We investigated different shapes in CLIP data and pronounce a filter step to distinguish different peak shapes and thus improve subsequent analysis tasks.

4.6 Computational Approaches to Posttranscriptional Gene Regulation in Human Biology and Disease

Katharina Zarnack (Goethe-Universität – Frankfurt am Main, DE)

Main reference Simon Braun, Mihaela Enculescu, Samarth T. Setty, Mariela Cortés-López, Bernardo P. de Almeida, F. X. Reymond Sutandy, Laura Schulz, Anke Busch, Markus Seiler, Stefanie Ebersberger, Nuno L. Barbosa-Morais, Stefan Legewie, Julian König, Kathi Zarnack:"Decoding a cancer-relevant splicing decision in the RON proto-oncogene using high-throughput mutagenesis". Nat Commun 9, 3315, 2018.
URL https://doi.org/10.1038/s41467-018-05748-7

We employ a systems approach to better understand how multiple protein complexes dynamically interact on pre-mRNA sequence to control splicing (splice code). As a prototypical example, we study the alternative splicing of the MSTR1 gene which is frequently altered in cancer. Starting with a high-throughput mutagenesis screen, the complex splicing patterns are interpreted using mathematical models to infer changes in the splicing kinetics and to identify causative mutations. Importantly, the measured effects correlate with RON alternative splicing in cancer patients bearing the same mutations. Moreover, they highlight the RNA-binding protein HNRNPH as a key regulator of RON splicing in healthy tissues and cancer. Our results thereby offer insights into splicing regulation and the impact of mutations on alternative splicing in cancer.

4.7 A Translational Repression Complex in Developing Mammalian Neural Stem Cells that Regulates Neuronal Specification

Kazan Hilal (Antalya International University, TR)

Joint work of Siraj K. Zahr, Guang Yang, Hilal Kazan, Michael J. Borrett, Scott A. Yuzwa, Anastassia Voronova, David R. Kaplan, Freda D. Miller
Main reference Siraj K. Zahr, Guang Yang, Hilal Kazan, Michael J. Borrett, Scott A. Yuzwa, Anastassia Voronova, David R. Kaplan, Freda D. Miller: "A Translational Repression Complex in Developing Mammalian Neural Stem Cells that Regulates Neuronal Specification", Neuron, Vol. 97(3), pp. 520–537.e6, 2018.
URL http://dx.doi.org/10.1016/j.neuron.2017.12.045

The mechanisms instructing genesis of neuronal sub- types from mammalian neural precursors are not well understood. To address this issue, we have characterized the transcriptional landscape of radial glial precursors (RPs) in the embryonic murine cortex. We show that individual RPs express mRNA, but not protein, for transcriptional specifiers of both deep and superficial layer cortical neurons. Some of these mRNAs, including the superficial versus deep layer neuron transcriptional regulators Brn1 and Tle4, are translationally repressed by their association with the RNA- binding protein Pumilio2 (Pum2) and the 4E-T protein. Disruption of these repressive complexes in RPs mid-neurogenesis by knocking down 4E-T or Pum2 causes aberrant co-expression of deep layer neuron specification proteins in newborn superficial layer neurons. Thus, cortical RPs are transcriptionally primed to generate diverse types of neurons, and a Pum2/4E-T complex represses translation of some of these neuronal identity mRNAs to ensure appropriate temporal specification of daughter neurons.

4.8 in vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by co-factors

Julian König (Institut für Molekulare Biologie – Mainz, DE)

Main reference F.X. Reymond Sutandy, Stefanie Ebersberger, Lu Huang, Anke Busch, Maximilian Bach, Hyun-Seo Kang, Jörg Fallmann, Daniel Maticzka, Rolf Backofen, Peter F. Stadler, Kathi Zarnack, Michael Sattler, Stefan Legewie, Julian König: "In vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by cofactors. Genome Res 28(5), 699–713, 2018.
URL https://doi.org/10.1101/gr.229757.117

Alternative splicing generates distinct mRNA isoforms and is crucial for proteome diversity in eukaryotes. The RNA-binding protein (RBP) U2AF2 is central to splicing decisions, as it recognizes 3' splice sites and recruits the spliceosome. We establish 'in vitro iCLIP' experiments, in which recombinant RBPs are incubated with long transcripts, to study how U2AF2 recognizes RNA sequences and how this is modulated by trans-acting RBPs. We measure U2AF2 affinities at hundreds of binding sites, and compare in vitro and in vivo binding landscapes by mathematical modeling. We find that trans-acting RBPs extensively regulate U2AF2 binding in vivo, including enhanced recruitment to 3' splice sites and clearance of introns. Using machine learning, we identify and experimentally validate novel trans-acting RBPs (including FUBP1, BRUNOL6 and PCBP1) that modulate U2AF2 binding and affect splicing outcomes. Our study offers a blueprint for the high-throughput characterization of in vitro mRNP assembly and in vivo splicing regulation.

4.9 Posttranscriptional regulation in cellular and time

Markus Landthaler (Max-Delbrück-Centrum – Berlin, DE)

Spatial compartmentalization of RNA is central to many biological processes and enables diverse regulatory schemes that exploit both coding as well as noncoding functions of the transcriptome. Spatiotemporal RNA dynamics are typically examined by single molecule imaging techniques, but can simultaneously only be applied to a small number of transcripts. The combination of metabolic RNA labeling with biochemical nucleoside transitions adds a broadly applicable temporal dimension to RNA sequencing. To obtain insights in the spatiotemporal mRNA distribution in mouse embryonic stem cells we are using SLAM-seq (in collaboration with the labs of Stefan Ameres and Nils Blüthgen), a method for time-resolved measurement of newly synthesized and pre-existing RNA in cultured cells, in combination with cellular fraction and biochemical isolations. Current efforts are aiming at measuring transcriptome-wide kinetics of mRNA export from the nucleus, association with membranes and ribosomes. The goal is to identify sequence and/or structure features that modulate the spatiotemporal distribution of mRNAs.

4.10 Deconstructing an essential RNA regulatory program

Donny Licatalosi (Case Western Reserve University – Cleveland, US)

Joint work of Leah L. Zagore, Molly M. Hannigan, Sebastian M. Weyn-Vanhentenryck, Raul Jobava, Maria Hatzoglou, Chaolin Zhang, Donny D. Licatalosi

The RNA binding protein DAZL is essential for gametogenesis, but its direct in vivo functions, RNA targets, and the molecular basis for germ cell loss in Dazl-null mice are unknown. Here, we mapped transcriptome-wide DAZL-RNA interactions in vivo, revealing DAZL binding to thousands of mRNAs via polyA-proximal 3' UTR interactions. In parallel, fluorescence-activated cell sorting and RNA-seq identified mRNAs sensitive to DAZL deletion in male germ cells. Despite binding a broad set of mRNAs, integrative analyses indicate that DAZL post-transcriptionally controls only a subset of its mRNA targets, namely those corresponding to a network of genes that are critical for germ cell proliferation and survival. In addition, we provide evidence that polyA sequences have key roles in specifying DAZL-RNA interactions across the transcriptome. Our results reveal a mechanism for DAZL-RNA binding and illustrate that DAZL functions as a master regulator of a post-transcriptional mRNA program essential for germ cell survival.

4.11 RNA-mediated transcriptional regulation: a systematic search for new players

Yael Mandel-Gutfreund (Technion – Haifa, IL)

Joint work of Amir Argoetti, Rina Ben-El, Shlomi Dvir, Dor Shalev , Nathan Salomonis, Yael Mandel-Gutfreund

Transcription factors (TFs) play a pivot role in embryonic stem cells as key pluripotent markers. In recent years, it has been shown that long non-coding RNAs (lncRNAs) are involved in activation and repression of pluripotency-related genes via epigenetic and transcriptional regulation. To predict novel interactions between TFs and lncRNAs with regulatory functions in pluripotency and differentiation, we sampled RNA from eleven time points during directed differentiation of human Induced Pluripotent Stem Cells (iPSs) to Cardiomyocytes. Analyzing the differential expression patterns of coding and non-coding RNAs across time revealed pairs of TFs and lncRNAs that are significantly co-expressed, suggesting co-regulatory relationships. To confirm direct interactions between TFs and lncRNAs we performed eCLIP (enhanced crosslinking and immunoprecipitation) followed by sequencing on selected TFs. Computational analysis of the CLIP data revealed a small subset of non-coding RNAs with significantly enriched protein binding peaks. Specifically, the eCLIP results signify a direct association between the STAT3 (signal transducer and activator of transcription 3) TF and the lncRNA NORAD (non-coding RNA activated by DNA damage) in human pluripotent cells. Strikingly, knockdown of NORAD in hESCs significantly impaired STAT3 localization to the nucleus. Based on our findings, we propose that lncRNAs may contribute to stemness by directly interacting with TFs, possibly acting as co-regulators to modulate and fine-tune the transcriptional program of their target genes.

4.12 A new method to predict novel trans RNA-RNA interactions

Irmtraud Meyer (Max-Delbrück-Centrum – Berlin, DE)

Joint work of Sabine Reißer, Irmtraud Meyer

Many key mechanisms of gene regulation happen on transcriptome level. Key examples include micRNA-mRNA interations, RNA editing and RNA splicing. These are already known to crucially determine the functional products of any given cell. At the core of these interactions are trans RNA-RNA interactions, i.e. direct interactions between two or more transcripts that may happen at different time points of the transcript's life in the cell. Compared to networks of protein-protein interactions, the universe of trans RNA-RNA interactions remains vastly underexplored. Even the most recent duplex-based experimental methods for probing the RNA interactome and RNA structurome in vivo have biases and inefficiencies. On the computational side, there already exist numerous prediction methods. These, however, either cater for specific classes of biological interactions (e.g. miRNA-mRNA) where the key features of the interaction site are already well-known or aim to predict novel classes of trans RNA-RNA interactions while having a range of significant limitations.

To overcome these challenges, we have developed a new computational method that can detect trans RNA-RNA interactions in an unbiased manner provided the corresponding

functional features have been conserved in evolution. Our method takes a given multiple-sequence alignment and a corresponding phylogenetic tree as input and predicts helices (i.e. a helix being defined as a stretch of consecutive base-pairs) that have been well conserved in evolution. Our method employs a fully probabilistic framework that compares for each candidate helix the likelihood of having evolved as based-paired entity to the likehilhood of having evolved as unpaired entity. The corresponding log-likelihood scores are derived from two probabilistic models of evolution that model how base-pairs evolve over time and how un-paired nucleotides evolve over time, respectively. Our method is capable of also estimate p-values for all predicted helices. Compared to two existing state-of-the-art programs, the prediction accuracy of our method is significantly higher for a test set of known snoRNA-rRNA interactions from Lai *et al.* [1]. Due to its time- and memory-efficiency, our method readily extends to long biological transcripts which has been one major limitation of existing methods. We thus hope to apply our method on transcriptome-wide data sets in order to identify novel biological classes of trans RNA-RNA interactions.

References

1 D. Lai, I. M. Meyer, A comprehensive comparison of general RNA-RNA interaction prediction methods, Nucleic Acids Research 44(7):e61 (2016)

4.13 Characterizing the snoRNome through transcriptomics profiling and RNA-RNA interactomics

Michelle Scott (University of Sherbrooke, CA)

SnoRNAs have long been characterized for their role as guides for the site-specific modification of rRNA. In recent years however, increasing numbers of studies report diverse novel functions for snoRNAs including the regulation of alternative splicing, the control of the stability of mRNAs and pre-mRNAs, the regulation of chromatin architecture and as essential intermediates in cell stress responses. The characterization of snoRNAs has lagged behind other main RNA families, likely in part due to the difficulty in quantifying them by RNA-seq, because of their strong structure. We have recently established a methodology to accurately measure the abundance of snoRNAs using a reverse transcriptase with low structure bias. Despite the assumed housekeeping role of snoRNAs, expression profiles across various normal human tissues show snoRNAs covering a wide abundance range and a subset displaying tissue specificity or tissue variability, which relates to their host gene and their targets. Our data show that approximately 65% of snoRNAs are uniformly expressed across all tissues considered. Uniformly expressed snoRNAs are typically highly expressed, are often encoded in translation-related protein-coding host genes, target ribosomal RNA and are strongly conserved across evolution. On the other hand, snoRNAs displaying variable expression across tissues are less expressed, less conserved, display characteristics of feedback relationships with their host genes and are more likely to be involved in non-canonical functions in post-transcriptional regulation such as the regulation of alternative splicing. We are also using machine learning approaches to predict canonical and non-canonical targets of snoRNAs and integrating expression profiling and target prediction to characterize the extent of snoRNA functionality.

4.14 RNA regulatory dynamics controlling human steroidogenesis

Neelanjan Mukherjee (University of Colorado – Aurora, US)

Joint work of Kimberly Wellman, Kent Riemondy, Austin Gillen, Amber Baldwin, Neelanjan Mukherjee

Human steroid hormones produced by the adrenal cortex control important physiology including metabolism, inflammation, blood pressure/volume, and sexual characteristics. Many human disorders are caused by the lack or excess of adrenal hormones. For example, 1 in 20 Americans suffer from high blood pressure caused by excessive adrenal aldosterone production. While the signaling components, transcriptional regulators, and steroidogenic enzymes necessary for production of hormones have been identified, little is known about post-transcriptional regulation of steroidogenesis by RNA-binding proteins (RBPs). Recently technological advances have revolutionized our ability to investigate RBP-driven RNA regulation, making it possible for the first time to investigate how these mechanisms contribute to steroidogenesis. We have recently carried out a screen for RBPs regulating human aldosterone production that revealed numerous RBPs. Many of these RBPs are regulators of cytoplasmic RNA stability and translation.

Prominent among the hits in our screen were members of the tristetraprolin (ZFP36) family of RNA-binding protein. These RBPs binds to AU-rich elements (ARE) in 3'UTRs and consequently destabilizes and/or translationally represses these ARE-containing mRNAs. Through time-course experiments we found that mRNA stability controls the temporal pattern of RNA expression during steroidogenesis. Indeed, mRNAs with AU-rich elements (AREs) in their 3' UTR were rapidly induced and cleared out in response to steroidogenic stimulation. Furthermore, depletion of either ZFP36L2 or ZFP36L1 significantly increased aldosterone levels. These represents one of the first RBPs implicated in control of human aldosterone synthesis. Notably, over-production of aldosterone is a major cause of hypertension, suggesting that failure of this negative feedback loop could have important implications for human health. Additionally, genome-wide association studies have reported variants in both ZFP36L1 and ZFP36L2 associated with changes in systolic blood pressure.

We propose the that the ZFP36 family of proteins operate a negative feedback loop that prevent overproduction of aldosterone by destabilizing and/or translationally repressing ARE-containing mRNAs encoding steroidogenic proteins. Our ongoing research will elucidate the mechanism underlying this negative feedback loop that controls aldosterone biosynthesis post-transcriptionally through the action of ZFP36L1/2 RNA binding. Post-transcriptional regulation of mRNA stability and translation by AREs is a critical gene regulatory pathway important in many different tissues and conditions. Finally, the adrenal cortex is amenable to the delivery of modified oligonucleotides; thus, our discoveries can facilitate the design of oligonucleotide therapeutics that can be used to precisely and specifically modulate steroidogenesis through RBP-RNA disruption.

4.15 Deep learning for protein-RNA interactions

Yaron Orenstein (Ben Gurion University – Beer Sheva, IL)

Joint work of Ilan Ben-Bassat, Benny Chor, Yaron Orenstein
Main reference Ilan Ben-Bassat, Benny Chor, Yaron Orenstein: "A deep neural network approach for learning intrinsic protein-RNA binding preferences", Bioinformatics, Vol. 34(17), pp. i638–i646, 2018.
URL http://dx.doi.org/10.1093/bioinformatics/bty600

Protein-RNA binding, mediated through both RNA sequence and structure, plays vital role in many cellular processes, including neurodegenerative-diseases. Modelling the sequence and structure binding preferences of an RNA-binding protein is a key computational challenge. Accurate models will enable prediction of new interactions and better understanding of the binding mechanism.

DLPRB is e a new deep learning based approach to learn RNA sequence and structure binding preferences from large biological datasets. DLPRB outperforms the state of the art, both in vitro and in vivo. Moreover, biological insights can be gained by applying neural networks to large datasets of protein-RNA interactions.

4.16 Dynamic post-transcriptional RNA regulation in early zebrafish development

Michal Rabani (The Hebrew University of Jerusalem, IL)

Joint work of Michal Rabani, Lindsey Pieper, Guo-Liang Chew, Alexander F. Schier
Main reference Michal Rabani, Lindsey Pieper, Guo-Liang Chew, Alexander F. Schier: "Massively parallel reporter assay of 3'UTR sequences identifies in vivo rules for mRNA degradation", Molecular Cell, Vol. 68(6), pp. 1083–1094, 2017.
URL https://doi.org/10.1016/j.molcel.2017.11.014

The stability of mRNAs is regulated by signals within their sequences, but a systematic and predictive understanding of the underlying sequence rules remains elusive. In this talk, I will introduce UTR-Seq, a combination of massively parallel reporter assays and regression models, to survey the dynamics of tens-of-thousands of 3'UTR sequences during early zebrafish embryogenesis. I will focus on the massive degradation of maternally provided mRNAs, a key developmental transition in early embryos that is shared in all animals, as a powerful system to study mRNA dynamics in the absence of de-novo transcription. Applying UTR-Seq in this system, we revealed two temporal degradation programs: a maternally encoded early-onset program and a late-onset program that accelerated degradation after zygotic genome activation. Our analysis identifies regulatory sequences with specific roles: stabilizing poly-U and UUAG signals and destabilizing GC-rich signals act via early-onset pathways; and miR-430 seeds, ARE signals and PUM sites promote late-onset degradation. These elements identified through UTR-Seq also influence the stability of endogenous maternal mRNAs. Finally, Sequence based regression models translated 3'UTR sequences into their unique decay patterns, and predicted the in vivo impact of sequence signals on mRNA stability. Their application led to the successful design of artificial 3'UTRs that conferred specific mRNA dynamics. By using UTR-Seq as a general strategy to uncover the rules of RNA cis-regulation, we aim to learn the code of genomic information within native maternal mRNAs that defines their unique decay profiles, and its physiological roles during early developmental transitions.

4.17 Exploring inter-domain cooperation in RNA binding proteins

Andres Ramos (University College London, GB)

Joint work of Giuseppe Nicastro, Robert Dagil, V. Castilla-Llorente, C. Gallagher, Ian A. Taylor, J. Ule, Andres Ramos

Main reference Robert Dagil, Neil J. Ball, Roksana W. Ogrodowicz, Fruzsina Hobor, Andrew G. Purkiss, Geoff Kelly, Stephen R. Martin, Ian A. Taylor, Andres Ramos: "IMP1 KH1 and KH2 domains create a structural platform with unique RNA recognition and re-modelling properties", Nucleic Acids Research, Vol. 47(8), pp. 4334–4348, 2019.

URL http://dx.doi.org/10.1093/nar/gkz136

Main reference Giuseppe Nicastro, Adela M. Candel, Michael Uhl, Alain Oregioni, David Hollingworth, Rolf Backofen, Stephen R. Martin, Andres Ramos: "Mechanism of β-actin mRNA Recognition by ZBP1", Cell Reports, Vol. 18(5), pp. 1187–1199, 2017.

URL http://dx.doi.org/10.1016/j.celrep.2016.12.091

Most RNA-binding proteins recognise their RNA targets with the combinatorial action of multiple RNA-binding domains. This complexity is tunable, and allows the target-dependent recognition of a diverse range of features and sequences, underlying the capability of individual proteins to regulate multiple steps of RNA metabolism. A key question in RNA regulation is how the domains cooperate in target recognition, both for individual targets and at the transcriptome level.

We answer this question on the IGF2-mRNA binding protein 1 (IMP1), a key regulator of RNA metabolism, transport and translation. IMP1 plays an important role in defining synaptic morphology in human neurons and has a general function in regulating cell motility and differentiation. Further, IMP1 is expressed at very low levels in most cells in the adult, but high level of IMP1 expression in cancer are connected to cancer cell invasion and to the final outcome of the pathology. At the molecular level, IMP1 regulates the localisation, translation and stability of different sets of mRNA targets. The protein contains six putative RNA-binding domains – two RNA recognition motifs (RRMs) and four K-homology (KH) domains – that are organized in two-domain units.

We are using an ensemble of in vitro (e.g. NMR, X-ray crystallography, BLI, CD etc) to characterise the interaction of the individual domains and the larger multi-domain units in RNA binding. We have then built computational models of the kinetic pathways followed by the individual binding events, and that we can relate to microscopy data. Our results indicate that the two domains within one unit are strongly coupled and that each KH di-domain unit can bind RNA with high affinity and re-model it. However, the sequence selectivity and binding mechanisms are different in the two di-domains, which act quasi-independently and with very different kinetic properties. Importantly, we find these concepts are overall valid at the transcriptome level by performing in a novel computational analysis to compare sets of in vivo transcriptome-wide binding data recorded on protein mutants where RNA binding of individual domains has been knocked off using structure driven mutations.

4.18 Coding regions regulate mRNA stabilities in human cells

Olivia Rissland (University of Colorado – Denver, US)

A new paradigm has emerged that coding regions can regulate mRNA stability. Here, due to differences in cognate tRNA abundance, synonymous codons are translated at different speeds, and slow codons then stimulate mRNA decay. To ask if this phenomenon also

occurs in humans, we isolated RNA stability effects due to coding regions with the human ORFeome collection. We find that coding regions change mRNA stability primarily through translation. Instability-associated codons are translated more slowly, providing the first connection in humans between elongation speed and mRNA decay. Surprisingly, and in contrast to the existing model, the encoded amino acid also plays a key role. Analysis of ribosome profiling datasets indicates that decoding rates generally determine elongation speeds and are themselves likely controlled by both tRNA abundance and charging. Thus, we propose that both codon and amino acid usage regulate human mRNA stability, which may allow for coordinated regulation of related genes.

4.19 Eukaryotic-wide reconstruction of RNA-binding protein specificity by joint matrix factorization

Alexander Sasse (University of Toronto, CA)

Joint work of Alexander Sasse, Debashish Ray, Kaitlin U. Laverty, Hong Zheng, Kate Nie, Mihai Albu, Matthew H. Weirauch, Timothy R. Hughes, Quaid Morris
Main reference Debashish Ray, Hilal Kazan, Kate B. Cook, Matthew T. Weirauch, Hamed S. Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, Hong Na, Manuel Irimia, Leah H. Matzat, Ryan K. Dale, Sarah A. Smith, Christopher A. Yarosh, Seth M. Kelly, Behnam Nabet, Desirea Mecenas, Weimin Li, Rakesh S. Laishram, Mei Qiao, Howard D. Lipshitz, Fabio Piano, Anita H. Corbett, Russ P. Carstens, Brendan J. Frey, Richard A. Anderson, Kristen W. Lynch, Luiz O. F. Penalva, Elissa P. Lei, Andrew G. Fraser, Benjamin J. Blencowe, Quaid Morris, Timothy R. Hughes: "A compendium of RNA-binding motifs for decoding gene regulation", Nature, Vol. 499(7457), pp. 172–177, 2013.
URL http://dx.doi.org/10.1038/nature12311

Messenger RNA (mRNA) maturation is defined by co- and post-transcriptional interactions with RNA binding proteins (RBPs). Most RBPs possess unique binding specificities towards sequence, or sequence-structure patterns, called motifs. The largest set of these sequence motifs has been measured by RNAcompete, an in vitro assay which measures binding strength of the protein to a designed pool of 250,000 RNA sequences (Ray et al. 2013). Previously, these measurements were used to infer motifs of uncharacterized eukaryotic RBPs that shared at least 70% sequence identity to a measured protein sequence. However, for sequences containing RNA recognition motif domains (RRM), the most abundant RNA binding domain, predictions based on sequence identity between 40 to 70% were ambiguous. To address this issue, we developed a new computational method, called joint matrix factorization (jMF), which infers binding preferences from peptide profiles (k-mers). JMF circumvents the need for sequence alignments, which can be error prone, and enables high confidence predictions for about 15% more RBPs, formerly ambiguous cases. Moreover, jMF predicts the importance of individual peptides for different binding specificities. Tested on co-complex structures of RRMs, these peptide scores showed significant improvements in predicting RNA binding sites from protein sequence compared to classical conservation scores. Combining RNAcompete measurements with computational predictions from JMF we increased the total number of eukaryotic RBPs with known specificities from 12,000 to 32.500 (RRM/KH 36%) across more than 700 species, leading to 121 motifs for Homo sapiens, 51 for C. elegans, and 48 for Arabidopsis thaliana. We used the inferred set of binding specificities for Arabidopsis thaliana and combined it with gene expression data from 67 tissue types to determine crucial post-transcriptional regulators. The extended compendium of measured and inferred RNA binding specificities will be available on the CisBP-RNA database (http://cisbp-rna.ccbr.utoronto.ca/)

4.20 Decoding regulatory protein-RNA interactions by combining integrative structural biology and large-scale approaches

Michael Sattler (Helmholtz Zentrum – München, DE)

Main reference Tim Schneider, Lee-Hsueh Hung, Masood Aziz, Anna Wilmen, Stephanie Thaum, Jacqueline Wagner, Robert Janowski, Simon Müller, Silke Schreiner, Peter Friedhoff, Stefan Hüttelmaier, Dierk Niessing, Michael Sattler, Andreas Schlundt, Albrecht Bindereif: "Combinatorial recognition of clustered RNA elements by the multidomain RNA-binding protein IMP3". Nature Communications 10:1–18, 2019
URL https://doi.org/10.1038/s41467-019-09769-8
Main reference Hamed Kooshapur, Nila Roy Choudhury, Bernd Simon, Max Mühlbauer, Alexander Jussupow, Noemi Fernandez, Alisha N. Jones, Andre Dallmann, Frank Gabel, Carlo Camilloni, Gracjan Michlewski, Javier F. Caceres, Michael Sattler: "Structural basis for terminal loop recognition and stimulation of pri-miRNA-18a processing by hnRNP A1". Nat Commun 9, 2479, 2018.
URL https://doi.org/10.1038/s41467-018-04871-9

RNA plays essential roles in virtually all aspects of gene regulation, where single-stranded or folded regulatory RNA motifs are recognized by RNA binding proteins (RBPs). Most eukaryotic RBPs are multi-domain proteins that comprise various structural domains to mediate protein-RNA or protein-protein interactions. Linked to this molecular mechanisms of the formation and function of regulatory protein-RNA complexes often involve dynamic structural ensembles and can be controlled by population shifts between inactive and inactive conformations. The domains in these proteins are often connected or flanked by intrinsically disordered regions, where posttranslational modifications can further modulate the protein-RNA interactions to regulate the biological activity. We employ integrative structural biology combining solution techniques such as NMR, small angle scattering (SAXS/SANS) and FRET with X-ray crystallography and biophysical techniques to unravel the molecular recognition and dynamics for the assembly and molecular function of regulatory RNP (ribonucleoprotein) complexes. Three examples are discussed that highlight the role of conformational changes and dynamics in the function of RNPs. 1) We found that the U2AF2 RNA binding specificity for Py-tract RNA depends on an intrinsically disordered linker region flanking the canonical RNA binding domains. 2) Recognition of multipartite cis-regulatory motifs by the multi-domain RBP IMP3 involves cooperative protein-RNA interactions. 3) Recognition of the terminal loop of the pri-miR-18a hairpin by hnRNP A1 involves partial melting of the upper stem region to enhance its processing and function. Our data provide unique insight into conformational dynamics underlying the regulation of essential biological processes. The combination of experimental biophysical and structural biology techniques with large-scale genome-wide mapping and efficient computational tools is essential to unravel the protein-RNA code.

4.21 GraphProt2: deep learning for graphs meets RBP binding site prediction

Michael Uhl (Universität Freiburg, DE)

CLIP-seq is the current state-of-the-art technique to experimentally determine transcriptome-wide binding sites of RNA-binding proteins (RBPs). However, since it relies on gene expression which is highly variable between conditions, it cannot provide a complete picture of the RBP

binding landscape. This necessitates the use of computational methods to predict missing binding sites, which is usually done by learning relevant features from identified sites and then use the learned model for prediction on unseen sequences. Here we present GraphProt2, a computational RBP binding site prediction method based on graph convolutional neural networks (GCNs). GraphProt2 converts the input binding sites into graphs and uses these for model training and prediction. In contrast to popular convolutional neural network (CNN) methods, this allows for variable length input as well as the possibility to add base pair information. Furthermore, additional features such as accessibility, conservation or region type information can be added as feature vectors to each node to improve predictive performance. Preliminary results show superior performance when compared to GraphProt as well as iDeepS, a CNN-based method that also utilizes a long short-term memory (LSTM) extension. For single RBP models, average accuracy in 10-fold cross validation over 33 eCLIP datasets was 86.13% (SD: +- 0.84) for GraphProt2, 81.87% (SD: +-1.20) for iDeepS, and 77.54% (no SD information given) for GraphProt.

4.22 Breaking apart 3'-UTRs to model in vivo post-transcriptional regulation

Charles E. Vejnar (Yale University – New Haven, US)

Main reference Charles E. Vejnar, Mario Abdel Messih, Carter M. Takacs, Valeria Yartseva, Panos Oikonomou, Romain Christiano, Marlon Stoeckius, Stephanie Lau, Miler T. Lee, Jean-Denis Beaudoin, Damir Musaev, Hiba Darwich-Codore, Tobias C. Walther, Saeed Tavazoie, Daniel Cifuentes, Antonio J. Giraldez: "Genome wide analysis of 3' UTR sequence elements and proteins regulating mRNA stability during maternal-to-zygotic transition in zebrafish", Genome Res., Vol. 29(7), pp. 1100–1114, 2019.
URL https://doi.org/10.1101/gr.245159.118

Post-transcriptional regulation plays a crucial role in shaping gene expression. During the Maternal-to-Zygotic Transition (MZT), thousands of maternal transcripts are regulated. However, how different cis-elements and trans-factors are integrated to determine mRNA stability remains poorly understood. Here, we show that most transcripts are under combinatorial regulation by multiple decay pathways during zebrafish MZT. Using a massively parallel reporter assay, we identified cis-regulatory sequences in the 3'-UTR, including U-rich motifs that are associated with increased mRNA stability. In contrast, miR-430 target sequences, UAUUUAUU AU-rich elements (ARE), CCUC and CUGC elements emerged as destabilizing motifs, with miR-430 and AREs causing mRNA deadenylation upon genome activation. We identified trans-factors by profiling RNA-protein interactions and found that poly(U) binding proteins are preferentially associated with 3'-UTR sequences and stabilizing motifs. We demonstrate that this activity is antagonized by C-rich motifs and correlated with protein binding. Finally, we integrated these regulatory motifs into a machine learning model that predicts reporter mRNA stability in vivo.

4.23 Deep Learning for Modeling Translation events

Jianyang Zeng (Tsinghua University – Beijing, CN)

Conventionally mRNAs are thought to only transfer the genetic information into protein sequences during translation. Recently more and more evidence has shown that mRNA sequences also encode the regulatory code that modulates translation initiation, elongation and termination. Now the high-throughput technique, Ribosome profiling, provides a large amount of data to measure the translational activities. In addition, deep learning has become a powerful machine learning tool for addressing the large-scale learning tasks. Then it remains unknown whether we can apply deep learning techniques to fully exploit the available large-ribosome profiling data to decode the sequence determents of translation regulation. Here, we develop three deep learning models to achieve this goal. In particular, we first apply a CNN model to predict the ribosome stalling events from the normalized ribosome footprints. Then we develop a deep reinforcement learning framework to select the most important codon features and make accurate prediction on ribosome density. Finally, we propose a hybrid deep learning model to predict translation initiation sites. Tests on real ribosome profiling data show that our models can achieve accurate predictions, outperform conventional learning models, and provide useful biological insights into understanding the translation mechanisms.

Participants

- Amir Argoetti
Technion – Haifa, IL
- Rolf Backofen
Universität Freiburg, DE
- Marina Chekulaeva
Max-Delbrück-Centrum – Berlin, DE
- Jörg Fallmann
Universität Leipzig, DE
- Jan Gorodkin
University of Copenhagen, DK
- Florian Heyl
Universität Freiburg, DE
- Eckhard Jankowsky
Case Western Reserve University – Cleveland, US
- Hilal Kazan
Antalya International University, TR
- Julian König
Institut für Molekulare Biologie – Mainz, DE
- Markus Landthaler
Max-Delbrück-Centrum – Berlin, DE
- Donny Licatalosi
Case Western Reserve University – Cleveland, US
- Yael Mandel-Gutfreund
Technion – Haifa, IL
- Irmtraud Meyer
Max-Delbrück-Centrum – Berlin, DE
- Neelanjan Mukherjee
University of Colorado – Aurora, US
- Uwe Ohler
Max-Delbrück-Centrum – Berlin, DE
- Yaron Orenstein
Ben Gurion University – Beer Sheva, IL
- Teresa Przytycka
National Center for Biotechnology – Bethesda, US
- Michal Rabani
The Hebrew University of Jerusalem, IL
- Andres Ramos
University College London, GB
- Olivia Rissland
University of Colorado – Denver, US
- Alexander Sasse
University of Toronto, CA
- Michael Sattler
Helmholtz Zentrum – München, DE
- Michelle Scott
University of Sherbrooke, CA
- Michael Uhl
Universität Freiburg, DE
- Charles E. Vejnar
Yale University – New Haven, US
- Katharina Zarnack
Goethe-Universität – Frankfurt am Main, DE
- Jianyang Zeng
Tsinghua University – Beijing, CN