Pacific Symposium on Biocomputing 14:545-548 (2009)

Post-transcriptional gene regulation: RNA-protein interactions, RNA processing, mRNA stability and localization

Co-chairs: Benjamin Blencowe[1,2], Steven Brenner[3], Timothy Hughes[1,2], Quaid Morris[1,2,4]

[1]Department of Molecular Genetics, University of Toronto, 1 King's College Rd., Toronto, ON, M5S 1A8, Canada
[2]Banting and Best Department of Medical Research, University of Toronto, 160 College Street, Toronto, ON, M5G 1L6, Canada
[2]Departments of Molecular and Cell Biology and Plant and Microbial Biology, University of California, Berkeley, California 94720, USA.
[4]Department of Computer Science, University of Toronto, 10 King's College Road, Toronto, ON, M5S 3G4, Canada

Background:

Regulation of gene expression at the post-transcriptional level, including the control of splicing, mRNA transcript stability, localization, and translation, is widespread in eukaryotes. An explosion of recent findings underscores both the predominance and complexity of PTR, including the discovery of ubiquitous but previously unknown regulatory mechanisms, such as microRNAs[1] (miRNAs); the observation that RNA recognition motif (RRM) and Kelch homology (KH) RNA-binding domains (RBDs) are among the most numerous protein domains in metazoan genomes, including the human genome[2, 3]; and the finding that, in several cases, RNA-binding proteins (RBPs) bind and presumably co-regulate sets of functionally-related transcripts[4]. An overall hypothesis has emerged that mRNA-containing ribonucleoprotein complexes (mRNPs) are the functional equivalent of "post-transcriptional operons"[5]. There are, however, relatively few examples of such "operons"; many putative PTR *cis*- and *trans*-regulatory factors were discovered by computational analysis of genome sequences, and have not been subjected to individual directed experimentation. Nonetheless, the sheer number of miRNAs (>500 in humans) and apparent sequence-specific RBPs (i.e. those carrying RRM, KH, and other RBDs) suggests that PTR mechanisms are as abundant as transcriptional regulatory mechanisms.

The study of PTR regulation presents a number of experimental challenges. Two of the major modes of PTR (control of subcellular localization and translation) cannot be assayed by analysis of total RNA extracted from cells; rather, they must be analyzed either *in situ* or via extraction of intact macromolecular structures (e.g. organelles, P-bodies, polysomes). Another major mode, transcript degradation rate, must be separated from transcript synthesis rate in order to make accurate measurements.

Coupled with these experimental challenges are computational challenges both in analysis of experimental data and in the modeling of *cis*-regulatory elements involved in PTR. For example, though alternative splicing can be assayed by analyzing total RNA using microarrays or next-generation sequencing, because alternatively spliced isoforms share most of their RNA sequence, estimating relative abundance levels from these data requires careful computational modeling [6-8]. Unlike DNA-binding proteins, many RBPs have both sequence and RNA secondary structure binding preferences – even among RRM and KH domains, which primarily recognize unpaired bases, roughly half of the proteins that have been characterized have a preference or requirement that the primary sequence is embedded in a particular type of secondary structure[3, 9, 10]. As

such, motif models of RBP binding preference should include both primary sequence and secondary structure features. However, most of the recent work in describing secondary-structure motifs has been focused on use of Covariance Models (CMs)[11, 12] to model families of non-coding RNA (ncRNA)[13, 14] using stochastic context-free grammars[15]. However, fitting CMs to collections of sequences bound by an RBP can be difficult because the procedures rely heavily on sequence alignment and compensatory mutations among paired bases[16] to detect features of RNA secondary structure. As such, it is not clear that these representations are appropriate for protein binding sites, which tend to be much smaller than ncRNAs and may be subject to different constraints. So, despite an increase variety of experiment methods available for identifying sequences bound by RNA-binding protein, either *in vivo* (e.g. RIP-chip[17] and CLIP[18, 19]) or *in vitro* (e.g., SELEX [20]), substantial effort is still needed to develop motif models that capture RBP binding preferences and can be easily learned from sequence data.

There also remain open questions in understanding how trans-acting regulators of PTR identify their targets. For example, though microRNA targeting has been extensively studied in the last few years, until recently, the most accurate predictors of miRNA target sites were conserved matches to the ~7bp seed region near the 5' end of the miRNA[21, 22]. Accessibility of target sites is now known to be an even more accurate predictor[23] though the target predictions of miRNAs are still far from complete and new features of target mRNA sequence continue to be discovered [24].

Summary

The goal of our workshop is to introduce some recent work in the area of post-transcriptional regulation to a wider computational community, discuss some of the unique computational problems faced in this area, and to present some preliminary solutions to these problems. In particular, we will focus on emerging computational and large-scale experimental strategies (e.g. microarray and deep sequencing) for investigating aspects of gene regulation at the post-transcriptional level, with an emphasis on the identification and characterization of the cis- and trans-acting RNA and protein components involved. We will also be exploring new developments in computational methods to detect and characterize cis-regulatory signals encoded in mRNAs.

Bibliography

1. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function**. *Cell* 2004, **116**(2):281-297.
2. Anantharaman V, Koonin EV, Aravind L: **Comparative genomics and evolution of proteins involved in RNA metabolism**. *Nucleic Acids Res* 2002, **30**(7):1427-1464.
3. Clery A, Blatter M, Allain FH: **RNA recognition motifs: boring? Not quite**. *Curr Opin Struct Biol* 2008, **18**(3):290-298.
4. Gerber AP, Herschlag D, Brown PO: **Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast**. *PLoS Biol* 2004, **2**(3):E79.
5. Keene JD: **RNA regulons: coordination of post-transcriptional events**. *Nat Rev Genet* 2007, **8**(7):533-543.
6. Anton MA, Gorostiaga D, Guruceaga E, Segura V, Carmona-Saez P, Pascual-Montano A, Pio R, Montuenga LM, Rubio A: **SPACE: an algorithm to predict and quantify alternatively spliced isoforms using microarrays**. *Genome Biol* 2008, **9**(2):R46.
7. Shai O, Morris QD, Blencowe BJ, Frey BJ: **Inferring global levels of alternative splicing isoforms using a generative model of microarray data**. *Bioinformatics* 2006, **22**(5):606-613.

8. Xing Y, Stoilov P, Kapur K, Han A, Jiang H, Shen S, Black DL, Wong WH: **MADS: a new and improved method for analysis of differential alternative splicing by exon-tiling microarrays**. *Rna* 2008, **14**(8):1470-1479.
9. Auweter SD, Oberstrass FC, Allain FH: **Sequence-specific binding of single-stranded RNA: is there a code for recognition?** *Nucleic Acids Res* 2006, **34**(17):4943-4959.
10. Valverde R, Edwards L, Regan L: **Structure and function of KH domains**. *Febs J* 2008, **275**(11):2712-2726.
11. Eddy SR, Durbin R: **RNA sequence analysis using covariance models**. *Nucleic Acids Res* 1994, **22**(11):2079-2088.
12. Sakakibara Y, Brown M, Hughey R, Mian IS, Sjolander K, Underwood RC, Haussler D: **Stochastic context-free grammars for tRNA modeling**. *Nucleic Acids Res* 1994, **22**(23):5112-5120.
13. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR: **Rfam: an RNA family database**. *Nucleic Acids Res* 2003, **31**(1):439-441.
14. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes**. *Nucleic Acids Res* 2005, **33**(Database issue):D121-124.
15. Yao Z, Weinberg Z, Ruzzo WL: **CMfinder--a covariance model based RNA motif finding algorithm**. *Bioinformatics* 2006, **22**(4):445-452.
16. Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D: **Identification and classification of conserved RNA secondary structures in the human genome**. *PLoS Comput Biol* 2006, **2**(4):e33.
17. Keene JD, Komisarow JM, Friedersdorf MB: **RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts**. *Nat Protoc* 2006, **1**(1):302-307.
18. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB: **CLIP identifies Nova-regulated RNA networks in the brain**. *Science* 2003, **302**(5648):1212-1215.
19. Ule J, Jensen K, Mele A, Darnell RB: **CLIP: a method for identifying protein-RNA interaction sites in living cells**. *Methods (San Diego, Calif* 2005, **37**(4):376-386.
20. Schneider D, Tuerk C, Gold L: **Selection of high affinity RNA ligands to the bacteriophage R17 coat protein**. *J Mol Biol* 1992, **228**(3):862-869.
21. Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M *et al*: **Combinatorial microRNA target predictions**. *Nat Genet* 2005, **37**(5):495-500.
22. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets**. *Cell* 2005, **120**(1):15-20.
23. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E: **The role of site accessibility in microRNA target recognition**. *Nat Genet* 2007, **39**(10):1278-1284.
24. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP: **MicroRNA targeting specificity in mammals: determinants beyond seed pairing**. *Mol Cell* 2007, **27**(1):91-105.