Marie-Sklodowska-Curie Actions

EvolSpliceKinetics

From co-transcriptional splicing kinetics to the evolutionary impact of exon and intron definition — EvolSpliceKinetics

Funding Agency European Commission

Funding Programme Horizon 2020

Call/Topic H2020-MSCA-IF-2018

Project reference 842695

Start date 01 Aug 2020

Duration 24 months

Total investment €147,815.04

Project Beneficiaries

Instituto de Medicina Molecular João Lobo Antunes (iMM), Portugal;

Researcher Rosina Savisaar

Supervisor Maria Carmo-Fonseca

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 842695

Summary

Rosina Savisaar

Biological systems often bewilder the observer with their staggering levels of complexity. This complexity creates a vulnerability – each of the cogs in the machine could potentially break, perhaps leading to disease. A prime example of such complexity is a process referred to as splicing, a key step in gene expression. A gene is said to be "expressed" when its DNA sequence is transcribed into an RNA molecule, which then either directly carries out the biological functions of the gene, or serves as a template for the production of a protein. In both cases, most RNAs must first undergo splicing – a processing step, where certain regions of the RNA ("introns") are removed and the remainder ("exons") ligated together again. Our genes are pock-marked throughout with scores of tiny sequence signals, which combine in a complex code, allowing the cell to recognize which regions are exons and which are introns. Disruption of either these sequence signals or the proteins that recognize them can lead to malformed RNAs being produced, sometimes with disastrous consequences. Indeed, about a third of disease-causing mutations in humans disrupt splicing.

Adding to this complexity, it is now known that often, introns located towards the start of the RNA molecule are spliced out whilst transcription of regions further down is still in progress. This is referred to as "co-transcriptional" splicing, and it opens up completely new ways of thinking about the process. Rather than simply considering the end product of splicing – which regions are removed and which ones remain – one can now turn the spotlight on the dynamics of the process. How do splicing and transcription affect one another, given that they often happen simultaneously? Are all introns removed equally fast? Do the sequence signals that control splicing differ for introns with slow and fast splicing?

In this project, I have studied the dynamics of co-transcriptional splicing in the fruit fly Drosophila melanogaster, a species whose genes have widely varying exon-intron structures. I analysed data from Native Elongating Transcript Sequencing (NET-seq), a method for sequencing RNAs that are still in the process of transcription ("nascent" RNAs). NET-seq captures nascent RNAs by targeting the enzyme that transcribes the genes into RNA, known as RNA Polymerase II (Pol II). From the nascent RNAs, one can verify whether the introns contained within have been spliced out. The splicing efficiency of each intron is then estimated using a proxy metric called the "splicing ratio" (SR). The higher the SR, the faster the intron is presumed to be spliced. The data can also be used to map the locations of Pol IIs and thus to infer the relative speed with which different gene regions were transcribed (although this may be confounded by Pol II phosphorylation patterns). This is important, as there is evidence for links between

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 842695

splicing and transcriptional pausing. However, these inferences can be marred by technical biases related to the nucleotide content of the sequences. I developed a simulation method to test our biological conclusions against such biases. I also designed an algorithm to determine which putative instances of Pol II pausing were the most reliable. This methodological work was disseminated through a blog post (https://imm.medicina.ulisboa.pt/news/the-peaks-and-valleys-of-the-nascent-transcriptome-in-drosophila-embryos), a seminar to students at the Faculty of Science of the University of Lisbon, and two practicals on an introductory bioinformatics course organized by the NGO Egypt Scholars, directed at students in Egypt.

I proceeded with a detailed characterisation of co-transcriptional splicing in the fruit fly, published as a co-first author paper in the RNA Journal (rna.078933.121v1), and presented at one national and three international conferences. SR varied drastically between introns, and correlated with properties of the intron in unexpected ways. Moreover, Pol II tended to pause at different locations depending on SR. I used Bayesian modelling to explore different hypotheses for mechanisms underlying these patterns. I concluded that the data could only be accounted for by a model where the same intron can stochastically switch between different modes of splicing kinetics.

Next, I checked whether the frequency or evolutionary conservation of splicing-related sequence signals depended on SR. I failed to uncover any significant patterns. This could be because the data was insufficient for such data-hungry analyses, or because variation in SR is either not functionally relevant or not controlled through sequence signals.

Carrying out the research described above led me to interact with scientists from a wide array of backgrounds. I realized how gravely research was often hampered by the fact that young researchers were not sufficiently trained in statistical thinking. Hence, a further goal of the project became to implement interventions to address this challenge. Three types of interventions were employed. Firstly, I designed an eight-week introductory statistics course. The course emphasized conceptual understanding, hands-on practice on real data and group work. I taught this course at the iMM in 2021, training a total of ca. 50 early career researchers. The course was repeated in the spring of 2022, as an online Arabic-language version, delivered in collaboration with Egypt Scholars. Both iterations of the course received overwhelmingly positive feedback.

Secondly, in the June of 2022, I organized an international summer school on applying modelling techniques to biological data. The summer school, funded by a Horizon 2020 grant, was attended by researchers from the iMM in Portugal, the Max-Delbrück Zentrum in Germany, the Weizmann Institute of Science in Israel, and the University of Oxford in the UK. The participants took part in five days of hands-on workshops, delivered by an international group of instructors.

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 842695

Thirdly, I worked individually with researchers to help them better understand their data. This included the supervision of 2 PhD students, 1 Master's student and one intern, as well as aiding several other researchers. This work has led to one co-first author publication (10.3390/biomedicines10020199), with at least three other manuscripts in preparation.

My current focus is to build on this training experience in order to reach even more researchers through both offline and online courses, and through individual consulting.

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 842695