Reaction-Diffusion Model of RNA-Splicing

December 3, 2016

Contents

1 Introduction

RNA-splicing is a process in which sections of precursor mRNAs, called introns, are removed to arrive at mature mRNA transcripts. Introns are defined at their upstream end by a 5' splice site and at their downstream end by a 3' splice site. At each of these locations there is a conserved motif, generally containing around 6-10 bits of information. RNA-splicing is observed across nearly all eukaryotes and it is believed that at least a third of all genetic diseases in humans are caused by disrupted splicing.

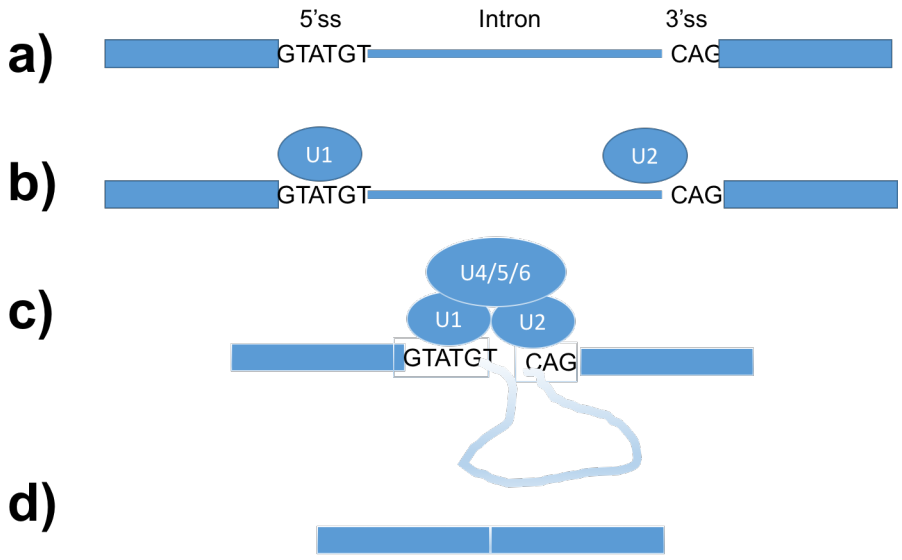

Figure 1: Splicing removes introns from precursor mRNAs (a) to arrive at mature transcripts (d). The splice sites are first recognized by U1 and U2 snRNP (b) followed by formation of a complex with the U4/5/6 tri-snRNP complex leading to splicing.

Splicing is orchestrated by thousands of proteins and RNAs, but the key players are U1 small nuclear ribo-protein (snRNP), which recognizes the 5'ss, U2 snRNP, which recognizes the 3' ss and the U4/5/6 tri-snRNP complex, which binds to RNA bound U1 and U2 to catalyze

splicing. I am leaving out branchpoint-binding protein, as well as the whole cast of splicing factors, enhancers and inhibitor for the sake of generality and simplicity.

In this paper, I implement a reaction-diffusion model of RNA splicing in order to gain insight into two key questions. First, introns can be very long with many human introns tens of thousands of nucleotides long. Even for more modestly sized introns it is unclear how the two ends are able to come together in order for the splicing reaction to procede. We will attempt to answer the question of if this can be explained by diffusion alone. Second, generally the 5'ss is upstream of the 3'ss, but this does not have to be the case. The splicing reaction could very easily proceed with a 3'ss upstream of the 5'ss resulting in a circular RNA. This phenomena is in fact observed and in recent years functional circular RNAs have been identified, but the vast majority of the time splicing proceeds with an upstream 5'ss. How can we explain this directionality?

2 Methods

In this section, I will describe our model of RNA splicing. An implementation of this model is publicly available on github (jpaggi/splicing-simulator). This can be broken generally into simulations of the snRNPs, the RNA, snRNP binding and splicing itself. Common to all was a discrete grid on which the molecules moved. The grid was set to be 50x50x50 with the edges looping.

2.1 Small nuclear ribo-protein simulation

SnRNPs were simulated as simply diffusing through the grid. On each time step, each molecule moved a step upwards, downwards, left, right, forwards or backwards. The molecules were all

initially placed at the origin, then given 300 iterations to diffuse prior to introduction of RNA. In the reported simulations we used 1000 U1 and U2 snRNPs and 5000 U3/4/5 tri-snRNPs.

2.2 RNA simulation

Our simulation of RNA is perhaps the most simplified and does not represent a physically realistic movement of RNA. On each time step, the RNA elongates by 1 (to simulate transcription) and is sent on a random flight from the origin. The random flight was done using a persistence length of 15 nucleotides, I.e. the molecule was discretized into 15 nucleotide segments and laid out in space tracing a random flight. While it is unsatisfying that this doesn't model the physical path that a RNA molecule will follow, it should capture the overall trends in its motion.

2.3 SnRNP binding

The snRNPs binding model is separated into two pieces, initial binding and release. On each time step, if a snRNP occupies the same grid cell as an unbound segment of RNA, it binds to it if a motif occurs that is more likely to be drawn from the corresponding position weight matrix (PWM), than from a random distribution. PWMs were computed for the 5'ss and 3'ss by summarizing the sequence at all annotated introns in budding yeast. On each time step, each bound snRNP releases with probability proportional to how much better the PWM explains the motif, than a random distribution. This scoring was done using the negative exponential of a standard bit score for the motif.

2.4 Splicing

Splicing itself occurs when a single grid cell is occupied by a segment of RNA bound to U1, a segment of RNA bound to U2 and a U4/5/6 tri-snRNP complex. When this occurs, we end the simulation and report the result.

3 Results

We ran our model on sequences from the S. cerevisiae, commonly known as budding yeast, genome. We chose budding yeast because they have particularly strong splice sites, short introns, and simple splicing patterns. Strong splice sites make it simpler to model the snRNP binding process. Short introns allowed us to simulate the reaction in a reasonable amount of time. Finally, budding yeast are believed to use a purely 'intron-definition' process for splicing (introns are recognized directly, as opposed to exons first being defined), which is intrinsically simpler to model.

We ran our model on all 273 intron containing genes and for each gene we ran 5 trials. We found that 56 introns were correctly predicted in at least one trial. Additionally, we found that 216 trials correctly detected the annotated 5'ss, while only 85 3'ss were correctly identified.

The intron lengths found in our simulations diverged significantly from the true distribution (Figure 2). Specifically, our predicted introns were shorter and some had 5'ss downstream from 3'ss, corresponding to negative distances. Out of the 1365 individual trials we found that 857 splices have upstream 5'ss, while, 322 do not.

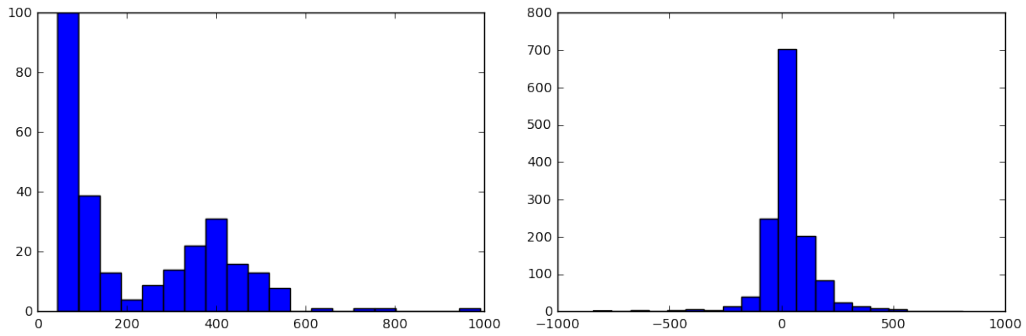

Figure 2: The true intron length distribution (left) differs substantially from the observed length distribution in our simulations (right)

4 Discussion

4.1 Entropic Adversities

While we were able to correctly simulate many short introns, we were able to predict very few long introns. While it could be the case that improved parameter tuning would allow us to perform better than we did, it seems unlikely that it will be able to fully explain how long intron ends are brought together. Human introns are often 10 to 100 times longer than yeast introns, so even if we could model the longest yeast introns, we would have far to go. If a simple diffusion model cannot explain splicing, then what are we missing?

First, our model of RNA is simplified in many areas, but one key area we are missing is RNA secondary structure. It has been shown that many long introns contain long step loops, which effectively reduce the length of the intron.

Second, our model leaves out the C-terminal domain tail (CTD tail) of the RNA polymerase. The CTD tail is known to bind snRNPs, favoring their placement onto the RNA. Perhaps, the CTD tail could additionally function to ”hold onto” snRNP bound snRNPs, in affect localizing

the splice sites.

4.2 Directionality of Splicing

Interestingly, our model did show a preference for splicing in the forward direction. This is interesting because at first glance our model treats 5'ss and 3'ss symmetrically. I believe that this trend can be explained by gene architecture. Examining the distribution of 5'ss positions, we find that they are nearly always at the extreme upstream end of the transcript. In this way, if the 5'ss is correctly identified, there is no option but to splice in the forward direction.

Could it be that introns have been selectively moved to the upstream end of yeast genes in order to enforce a forward direction of splicing? Another possible explanation for this trend is that introns hold transcriptional enhancers, which have evolved to be closer to the transcription start site. However, nearly all introns are very close to the beginning of the transcript, but it seems unlikely that all of these contain enhancers.

4.3 Moving Forward

There is a lot of room to improve our simulations. Perhaps, the most pressing concern is the model of RNA mechanics. Our current model does not model a true motion of RNA and ignores important characteristics of RNA such as secondary structure. Additionally, it is possible that snRNPs have more sophisticated mechanisms for searching for splice sites, such as 1D diffusion down RNA molecules. Also, our model for snRNP unbinding is based on the motifs observed for completed splicing events, not the intrinsic prefernces of the snRNPs.

Finally, splicing is an incredibly complicated process. In order to accurately model it would involve modeling a hundred fold more proteins at a much finer level of detail than shown here. Ideally work would be done to characterize each component more completely.