Contribution ID: **12**   Type: **Talk**

Machine Learning to Identify Gene Structure

Saturday 18 February 2023 13:45 (15 minutes)

The human genome is divided into exons and introns. Exons are the expressed portion of the gene and must be separated from the introns (the "junk"portion of DNA) during a process called splicing before they can be expressed as proteins. This division is import for alternate splicing, where the same gene can be expressed with different composite exons. But this can result in genes from two different locations fusing: this is a leading cause of cancer. Properly identifying exons and introns is an important step in studying gene structure and its role in cancer. Comparison of cancer fusion junctions (the location where two genes fuse) show a high prevalence of sequence similarity between exons of one gene and introns of its fusion. To measure this, we compared and scored sequences at the fusion junction, weighting their importance based on position from the fusion junction of the fusion genes to their expected sequence, what should be there if the genes were properly transcribed and spliced. While this suggests a splicing error it does not illuminate the underlying mechanism. For this reason, machine learning is being implemented to look for further structure. A convolution layer is added to identify exons, introns, and untranslated regions (a third region of the gene that plays a role in translation into proteins).

Academic year

3rd year

Research Advisor

Gemunu Gunaratne

Primary author: SPEAKMAN, Ethan

Presenter: SPEAKMAN, Ethan

Session Classification: Parallel Session 2

Track Classification: Biological and Statistical Physics