Recitation 7

Gene Splicing, Dimensionality Reduction

Corban Swain

MIT - 6.802 / 6.874 / 20.390 / 20.490 / HST.506 - Spring 2020

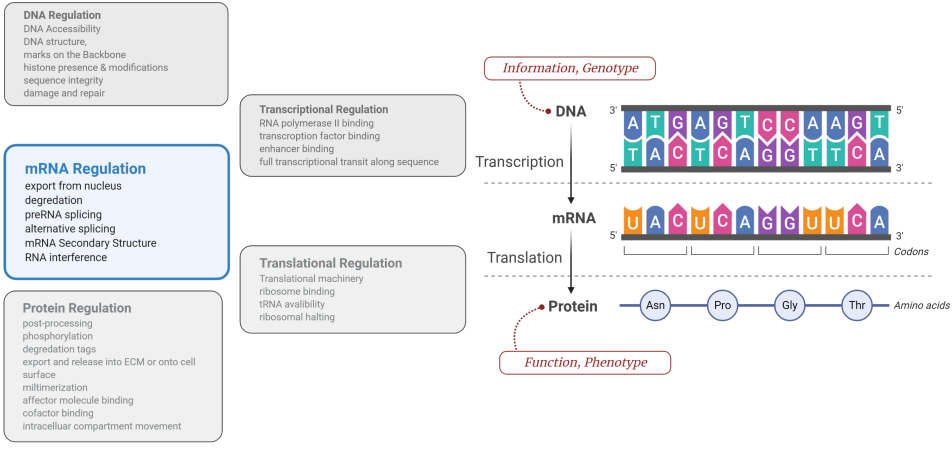
2020-04-02 / 2020-04-03

Outline

- RNA splicing and splicing codes
- Upsampling in the context of gene expression measurements
- Principle component Analysis Worked Example
- t-SNE “Perplexity” Meaning & Calculation

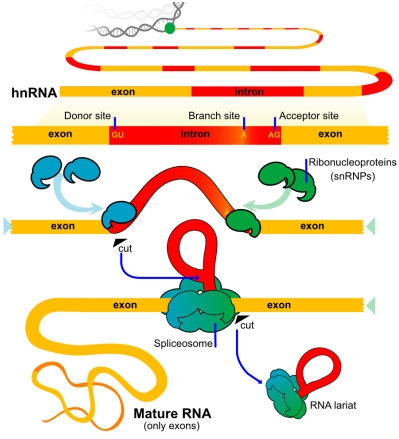
Gene Splicing Overview I

RNA Splicing as an element of the regulatory mechanisms that affect the mRNA molecule, the second key molecule of the central dogma.



Gene Splicing Overview II

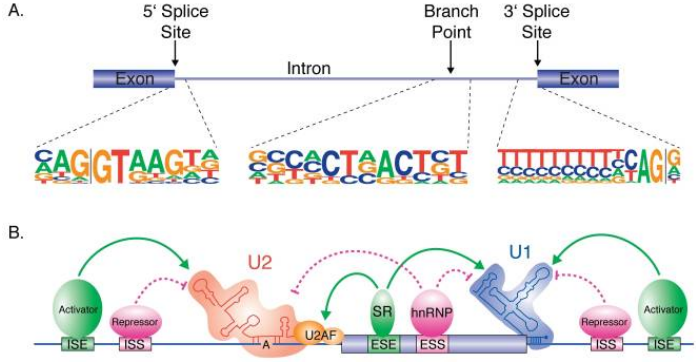
For splicing to occur (1) a number of protein components known as the "splicosome" must bind to the mRNA and (2) the prePRNA must adopt a specific secondary structure which brings the 3' end of the first exon in proximity to the 5' end of the following exon.



Gene Splicing Overview III

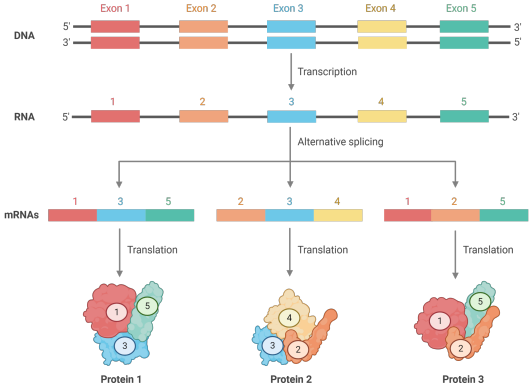
There are many regulatory elements which can control how splicing events occur, including:

- Consensus Sequence for 5' splice site, branch point, and 3' splice site
- pre-RNA Secondary Structure
- Intronic Splicing Enhancers & Silencers
- Exonic Splicing Enhancers & Silencers
- Splicesomal Component Binding and Reaction Catalysis



Gene Splicing Overview IV

Furthermore, these regulatory elements make it possible for a single preRNA transcript to yeild many different mRNA products and therefore different protein products.



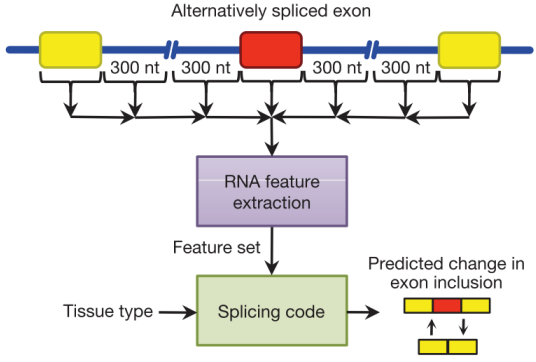
Dysregulation of these alternative splicing outcomes is contributor protein dysfunction and trascriptiome instability, contributing to pathologies including cancer and drug addiction.

Lec. 14 Paper: "Deciphering the splicing code" I

In their 2010 paper, Barash *et al.* attempt to predict how different splicing events occur in a tissue specif manner by using the surrounding sequence context.

- **The Problem:** Being able to predict what splicing variants will occur in a given context would make it possible gain a deeper understand and give scients the ability to predict and design the processing of RNA transcripts. No robust splicing code exists.
- **The Goal:** Infer the regulatory splicing code from
- **The Method:** A given preRNA transcript is divided into 300 nucleotide windows. Each window is evaluated for a number of utilize "known motifs" which were expected to affect splicing, "new motifs" which were not reported to affect splicing, "short motifs" 1 - 3 nt long, "structural features" which were known to affect secondary structure, among many other features. To develop their splicing code, they selectively pick feature from this "compendium" and use them at differnt thresholds to predict the inclusion or exclusion of a given exon.

Lec. 14 Paper: "Deciphering the splicing code" II

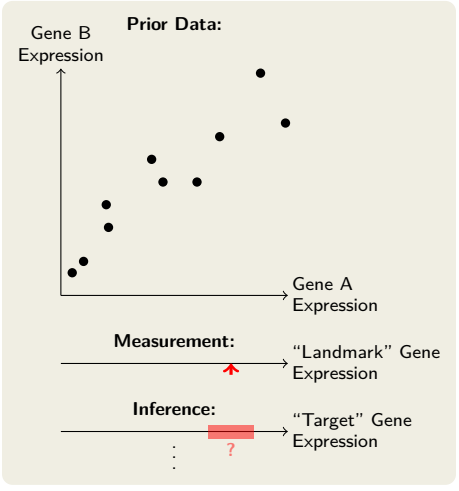


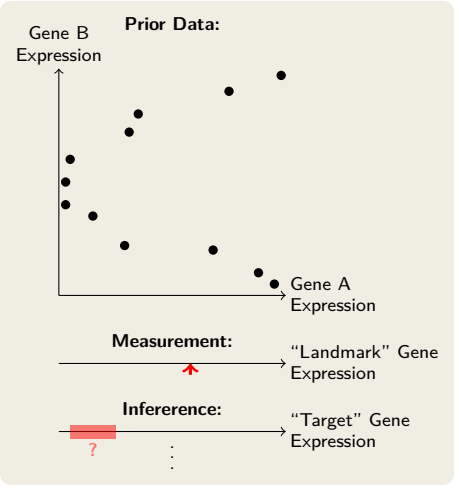
- **The Results:** They are able to identify tissue-specific splicing mechanisms and predict new mechanisms for developmental regulation.

Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., ... Frey, B. J. (2010). Deciphering the splicing code. *Nature*, 465(7294), 53–59. https://doi.org/10.1038/nature09000

Predicting Many Genes from Few Genes

To put the idea of "upsampling" in a gene expression context, let's take two simple examples:





Expand these examples many dimensions (*hundreds of landmark genes; tens of thousands of infered genes*) for some context of what expression upscaling approaches are hoping to solve.

Lec. 14 Paper: "Gene expression inference with deep learning" I

In their 2016 paper, Yifei Chen *et al.* use a deep learning approach upsample a select group of 1000 genes to infer whole geneome expression profiles.

- **The Problem:** Whole geneome expression profilling is too expensive to be generally applied in academic lab settings when there are many different conditions to be measured. Existing methods at the time (NIH LINCS) relied on linear regression for predection of 9.1k different target genes from a selection of ≈950 landmark genes.
- **The Goal:** Leverage a deep learning approach to capture non-linear relationships between the landmark and target genes and there fore make predictions with lower error.
- **The Method:** Their neural-network approach (D-GEX) is a feed forward network with either 1, 2, or 3 fully-connected hidden layers (not very "deep") containing 3k-9k hiddne units. They leverage dropout for model regularization and momentum methods for gradient descent. Control methods were linear regressions and k-nearest neighbors (KNN). For the KNN method, the k-nearest landmark genes are determined for each target geny based on Euclidian distance, at testtime a prediction is made for each target gene by taking the average expression of it's k landmark genes.

Lec. 14 Paper: "Gene expression inference with deep learning" II

- **The Results:** The neural network approach predicted target genes with less error than the linear and KNN approaches. Accuracy for the best method, NN with 3 hidden layers, was on the order of 70%.

Table 1. The overall errors of LR, LR-L1, LR-L2, KNN-GE and D-GEX-10% with different architectures on GEO-te

	Number of hidden units		
	3000	6000	9000
Number of hidden layers			
1	0.3421 ± 0.0858	0.3337 ± 0.0869	0.3300 ± 0.0874
2	0.3377 ± 0.0854	0.3280 ± 0.0869	0.3224 ± 0.0879
3	0.3362 ± 0.0850	0.3252 ± 0.0868	**0.3204** ± **0.0879**
LR		0.3784 ± 0.0851	
LR-L1		0.3782 ± 0.0844	
LR-L2		0.3784 ± 0.0851	
KNN-GE		0.5866 ± 0.0698	

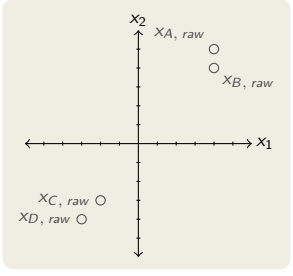
Chen, Y., Li, Y., Narayan, R., Subramanian, A., & Xie, X. (2016). Gene expression inference with deep learning. *Bioinformatics*, 32(12), 1832–1839. https://doi.org/10.1093/bioinformatics/btw074

PCA: Worked Example I

Consider that we have a set of 4 points

$$\mathbf{X}_{\mathbf{raw}} = \begin{bmatrix} 4 & 5 \\ 4 & 4 \\ -2 & -3 \\ -3 & -4 \end{bmatrix}$$

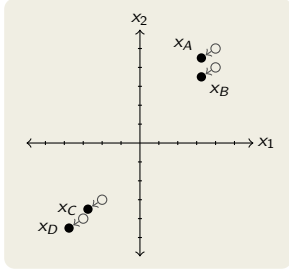
where each point has 2 dimensions; our goal will be to project these points onto the dimsnsion of highest variance (i.e. the first prinxiple component axis).



PCA: Worked Example II

1. We'll first need to shift the points to be zero centered along each dimension.

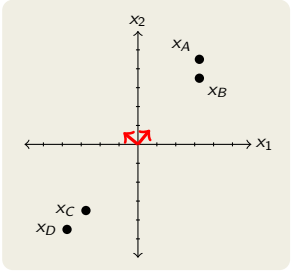
$$\mathbf{X} = \mathbf{X}_{\mathbf{raw}} - \overline{\mathbf{X}_{\mathbf{raw}}} = \begin{bmatrix} 3.25 & 4.5 \\ 3.25 & 3.5 \\ -2.75 & -3.5 \\ -3.75 & -4.5 \end{bmatrix}$$



PCA: Worked Example III

2. We then compute $\mathbf{X}^\mathsf{T}\mathbf{X}$ and find it's eigenvalues and eigenvectors.

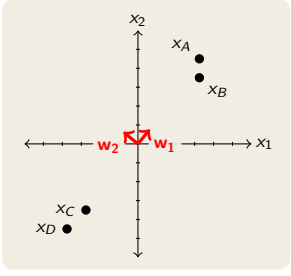
$$\mathbf{X}^\mathsf{T}\mathbf{X} = \begin{bmatrix} 42.75 & 52.5 \\ 52.5 & 65.0 \end{bmatrix} \quad \Rightarrow \quad \mathbf{Q} = \begin{bmatrix} -0.77 & 0.63 \\ 0.63 & 0.77 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 0.1 & 0 \\ 0 & 35.8 \end{bmatrix}$$



PCA: Worked Example IV

3. We now sort our eigenvectors by the eigenvalues, from largest (first principal components) to smallest (last principal components).

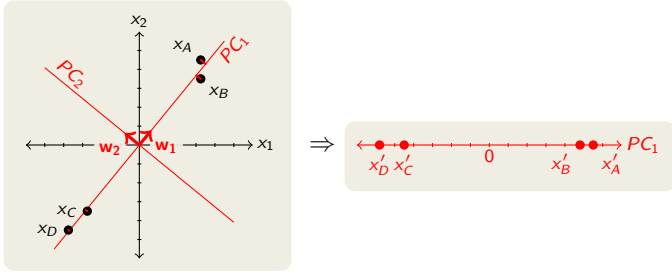
$$\mathbf{w_1} = \begin{bmatrix} 0.63 \\ 0.77 \end{bmatrix}, \quad \mathbf{w_2} = \begin{bmatrix} -0.77 \\ 0.63 \end{bmatrix}$$



PCA: Worked Example V

4. We can use $\mathbf{w_1}$ to project the 4 points into "PC1" space.

$$\mathbf{X_{PC1}} = \mathbf{X}\mathbf{w_1} = \begin{bmatrix} 5.5 \\ 4.8 \\ -4.5 \\ -5.8 \end{bmatrix}$$



Next Recitation

- Single-cell RNA Sequencing (scRNA-Seq)
- scRNA Seq Batch Correction
- Deep learning methods for scRNA
- Genetics Overview
- tSNE Perplexity Worked Example (from Recitation 6)