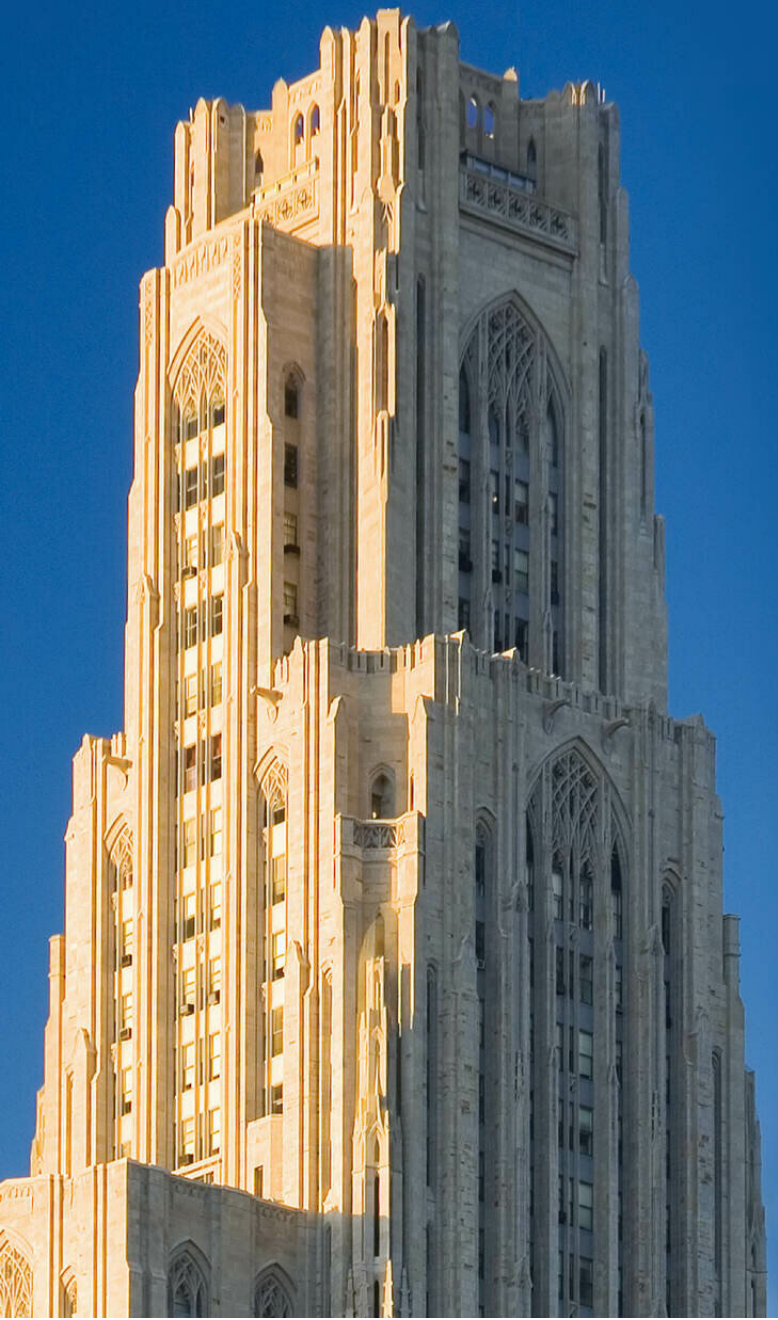


Computational Biology Seminar (BIOSC 1630)

Class 02 - September 06, 2023

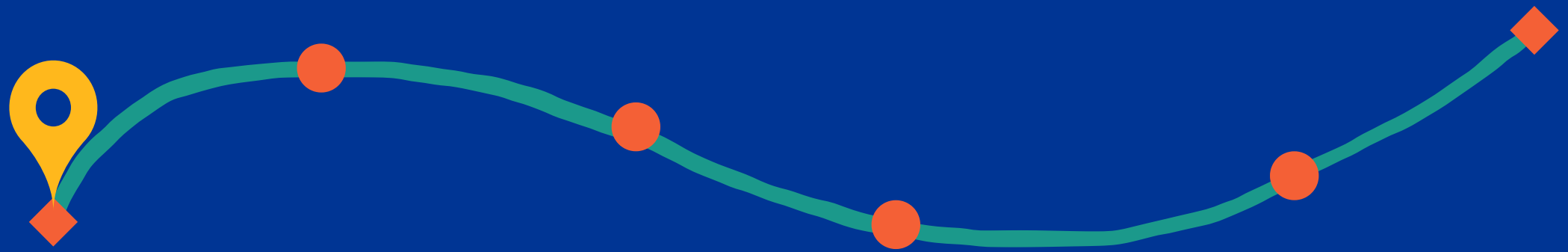
University of Pittsburgh

Announcements

- Drop-in discussions will be on Thursdays from 11 AM - 12 PM in 315 Clapp Hall
- Please fill out this survey to help me schedule your presentation
- No feedback in suggestion box
- I moved up the paper proposal to this week

Today's focus

- Use skimming and scanning techniques.
- Apply annotation, highlighting, and note-taking methods.
- Identify research questions, hypotheses, and objectives.
- Evaluate the significance and relevance of scientific studies.
- Create concise summaries by synthesizing key information.



We are here: Speed-reading techniques

Information overload

Around how many science and engineering papers were published per day in 2018?

Top Hat **7,123**!

We must be judicious in what articles we spend our time reading.

Speed reading

Subvocalization

When you are reading this sentence, is your inner monolouge saying them?

That is subvocalization

Some people say . . .

If you want to read fast, do not do this

If you want to comprehend, probably does not hurt

Speed reading

Scanning

The "Where's Waldo?" searching method for keywords

Fun fact: In other cultures, it is not "Waldo" but "Wally", "Willy", and others

Scan these paragraphs in a few seconds and tell us some keywords that are important

My selections

Motivation: Long-read RNA sequencing technologies are establishing themselves as the primary techniques to detect novel isoforms, and many such analyses are dependent on read alignments. However, the error rate and sequencing length of the reads create new challenges for accurately aligning them, particularly around small exons.
Results: We present an alignment method uLTRA for long RNA sequencing reads based on a novel two-pass collinear chaining algorithm. We show that uLTRA produces higher accuracy over state-of-the-art aligners with substantially higher accuracy for small exons on simulated and synthetic data. On simulated data, uLTRA achieves an accuracy of about 60% for exons of length 10 nucleotides or smaller and close to 90% accuracy for exons of length between 11 and 20 nucleotides. On biological data where true read location is unknown, we show several examples where uLTRA aligns to known and novel isoforms containing small exons that are not detected with other aligners. While uLTRA obtains its accuracy using annotations, it can also be used as a wrapper around minimap2 to align reads outside annotated regions.
Availabilityand implementation: uLTRA is available at https://github.com/ksahlin/ultra.
Contact: ksahlin@math.su.se
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The transcriptome has been identified as an important link between DNA and phenotype and is therefore analyzed in various biological and biomedical studies. For these analyses, RNA sequencing has established itself as the primary experimental method. Some of the most common transcriptome analyses using RNA sequencing data include predicting and detecting isoforms and quantifying their abundance in the sample. These analyses are fundamentally underpinned by the alignment of reads to genomes. As a transcriptomic read can contain multiple exons, alignment algorithms are required to handle split alignment of a read to multiple exonic regions of the genome, referred to as a spliced alignment.

Spliced alignment is a challenging computational problem, and a plethora of different alignment algorithms have been proposed for splice alignment of short-read RNA-seq, with some of the key algorithmic advances given in TopHat (Trapnell *et al.*, 2009), STAR (Dobin *et al.*, 2013), HISAT (Kim *et al.*, 2015), GMAP (Wu *et al.*, 2016) and HISAT2 (Kim *et al.*, 2019). While short-read RNA sequencing has shown unprecedented insights into transcriptional complexities of various organisms, the read-length makes it difficult to detect isoforms with complicated splicing structure and limits quantification of isoform abundance (Zhang *et al.*, 2017).

Long-read transcriptome sequencing protocols such as Pacific Biosciences (PacBio) Iso-Seq sequencing (Wang *et al.*, 2016) and Oxford Nanopore Technologies (ONT) cDNA and direct RNA sequencing (Workman *et al.*, 2019) are now establishing themselves as the primary sequencing techniques to detect novel isoforms. Long-read sequencing technologies can sequence transcripts from end to end, providing the full isoform structure and therefore offer accurate isoform detection and quantification. Such protocols have opened up the possibility to investigate the isoform landscape for genes with multiple gene copies (Sahlin *et al.*, 2018) and complex splicing patterns (Tseng *et al.*, 2019), as well as to accurately decipher alleles (Tilgner *et al.*, 2014) and cell-specific (Gupta *et al.*, 2018) isoforms. However, the long-read technologies also offer new algorithmic challenges because of the higher error rate and longer sequencing length which makes most short-read alignment algorithms unsuitable for long-read splice alignment (Križanović *et al.*, 2018). Therefore, long transcriptomic reads have, similarly to short reads, prompted splice alignment algorithm development. Some short-read aligners have been modified for long-read splice alignment (Dobin *et al.*, 2013; Wu *et al.*, 2016), while other aligners have been designed for splice alignment of long reads (Boratyn *et al.*, 2019; Li, 2018; Liu et al., 2019; Marić *et al.*, 2019). A recent method also suggested improving long-read splice alignments using ensemble prediction of splice sites (Parker *et al.*, 2021). First,

Sahlin, K., & Mäkinen, V. (2021). Accurate spliced alignment of long RNA sequencing reads. *Bioinformatics, 37*(24), 4643-4651. DOI: 10.1093/bioinformatics/btab540.

Speed reading

Skimming

Quickly reading for key points with a lot of skipping around

Beginning and end of paragraphs are hotspots

My selections

Sahlin, K., & Mäkinen, V. (2021). Accurate spliced alignment of long RNA sequencing reads. *Bioinformatics, 37*(24), 4643-4651. DOI: 10.1093/bioinformatics/btab540.

BIOSC1630

3.3 Splice site annotation performance on SIRV

While simulated data is good for comparisons due to the availability of ground truth annotations, it does not fully capture the error profiles present in sequencing data. We used a subset of 59 isoforms with distinct splice site positions from the ONT cDNA SIRV dataset (Sahlin and Medvedev, 2021) to investigate alignment performance around splice sites (for details see Supplementary Note SC). In this dataset we have a complete isoform annotation and the sequenced isoforms are known. We observed that uLTRA was able to align more reads to the isoforms, particularly to one isoform that contains an 8 nt long exon. deSALT and minimap2 did not align the large majority of reads that contained the exon (Supplementary Fig. S4E). Overall, uLTRA's alignments were more equally distributed across the 59 isoforms, as is expected in the SIRV E0 mix (see Supplementary Note SC). More details about the analysis and results are described in Supplementary Note SC.

3.4 Biological data

We also used an Alzheimer brain Iso-Seq dataset (denoted ALZ) and an ONT cDNA sequencing dataset from Drosophila (Sahlin and Medvedev, 2021) (denoted DROS). Both datasets have been processed with respective bioinformatics pipelines to select only the reads containing full-length transcripts (for details see Supplementary Note SD).

We neither have the correct read annotations, nor are we guaranteed to have a complete gene annotation for the biological datasets, which presents a challenge when evaluating accuracy. We took the following approaches. We first compared the alignment algorithms according to the alignment categories defined in Tardaguila *et al.* (2018) (presented in the next section). Secondly, we looked at the alignment concordance between methods. Here, we investigated concordance with respect to both alignment location on the genome and concordance based on the alignments around exons. Thirdly, we provide several examples of uniquely detected isoforms by uLTRA (and uLTRA_mm2), which demonstrate the caveats with alignment concordance analysis without ground truth.

3.4.1 Alignment categories on biological data

We classified alignments using the categories defined in Tardaguila *et al.* (2018). As in Tardaguila *et al.* (2018), we classify an alignment of a read to the genome as a Full Splice Match (FSM), Incomplete Splice Match (ISM), Novel In Catalog (NIC), Novel Not in Catalog (NNC) or NO_SPLICE. An FSM alignment means that the combination of splice junctions in the read alignment has been observed and annotated as an isoform. An ISM alignment means that the combination of splice junctions is in the annotation, but it is missing junctions compared to the annotated models in either the 3′ or 5′ end. A NIC alignment consists of junctions that all appear in the annotation, but not together in a single isoform. An NNC alignment means that the read aligns with at least one junction that is not in the annotation, while NO_SPLICE are all alignments without splice sites. These alignment categories are important for various downstream isoform detection methods such as SQANTI (Tardaguila *et al.*, 2018), TAMA (Kuo *et al.*, 2020) or TALON (Wyman *et al.*, 2019). See Tardaguila *et al.* (2018) for details regarding these definitions.

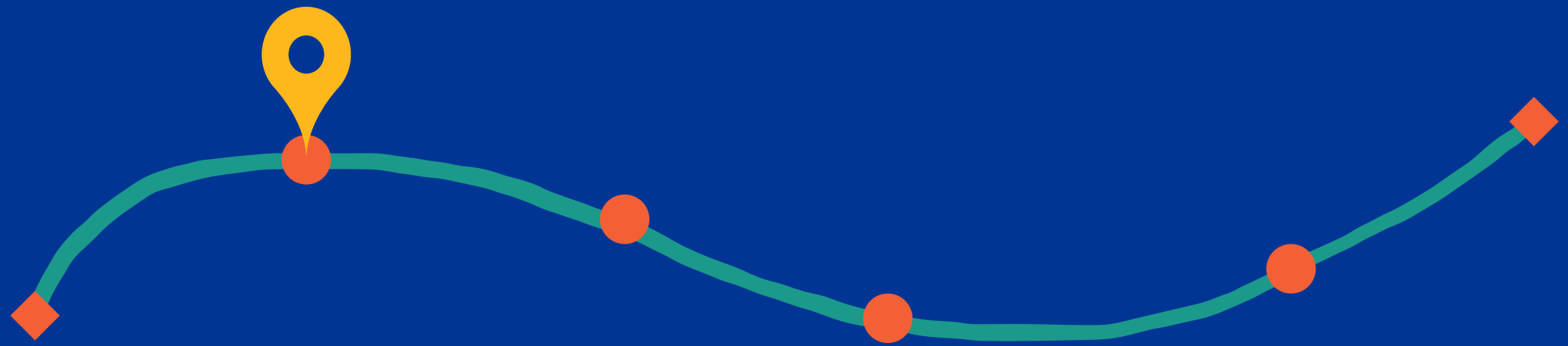
Overall, the aligners and their different modes produce a similar distribution of the different alignment categories on both the DROS (Fig. 3A) and ALZ (Fig. 3B) datasets. We observe that uLTRA, uLTRA_mm2 and deSALT_GTF align more FSM reads than deSALT, minimap2 and minimap2_GTF. In the ALZ dataset, uLTRA has many unaligned reads due to a large fraction of reads (17.6%) aligning outside uLTRA indexed regions. This highlights the benefit of not being limited to alignments around gene regions when aligning transcriptomic data even for well annotated genomes. We observe that the other aligners, including uLTRA_mm2, have no unaligned reads. Instead, they attribute a larger fraction of reads in the category NO_SPLICE (Fig. 3B). It is known that a substantial fraction of reads in long-read transcriptome sequencing data is coming from so-called intra-priming reads (Tardaguila *et al.*, 2018). These reads are characterized by aligning without splice junctions to an unannotated genome location that contains a poly-A stretch downstream from their 3′ end. While not fully characterized, these reads are likely to be artifacts in the sequencing protocol and often filtered out in downstream analysis (Tardaguila *et al.*, 2018).

We further investigated concordance in alignments within the different categories between uLTRA_mm2, deSALT_GTF and minimap2. They represent the best setting for each aligner, respectively, based on our accuracy evaluation on simulated data (Fig. 2, Supplementary Figs S1 and S2) and alignment consistency analysis on the synthetic SIRV data (Supplementary Fig. S4).

3.4.2 Alignment concordance on biological data

We looked at alignment concordance both with respect to genomic region (globally) and around exons (locally). A detailed description is found in Supplementary Note SE. Overall, we observed that 90.3% and 98.6% of all aligned reads had globally concordant alignments in DROS and ALZ, respectively (Supplementary Fig. S5). This indicates that the mapping region is largely consistent between aligners and that most of the variability occurs in alignments around exons. We also report local alignment concordance for each category, which was lower across each category (Supplementary Figs S6 and S7).

We also looked in more depth at the concordance of unique isoforms detected in the data that had FSM predictions (Supplementary Fig. S8). In total, 93.6% and 90.1% of the total unique isoforms with FSM alignments were aligned to by all the three methods on



We are here: Reading for gist

Reading for gist

Activity time!

We are here: Active reading techniques

Reading for comprehension

Understanding scientific literature requires active engagement

For example,

Highlight and annotate

Mark important sentences or write small notes in the margin

Use sparingly! This is more for when you need to go back and review

Pause and reflect

Write down what you understood about what you just read

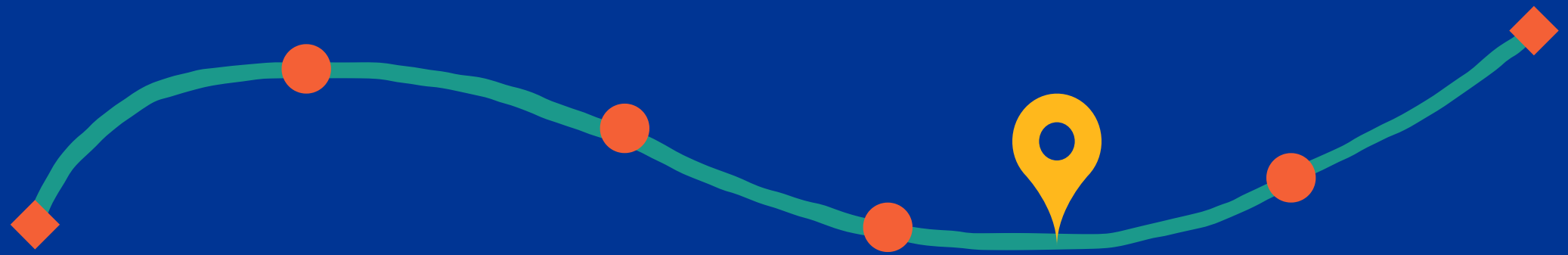
Explain to someone or something

Teaching is the best way to test your understanding

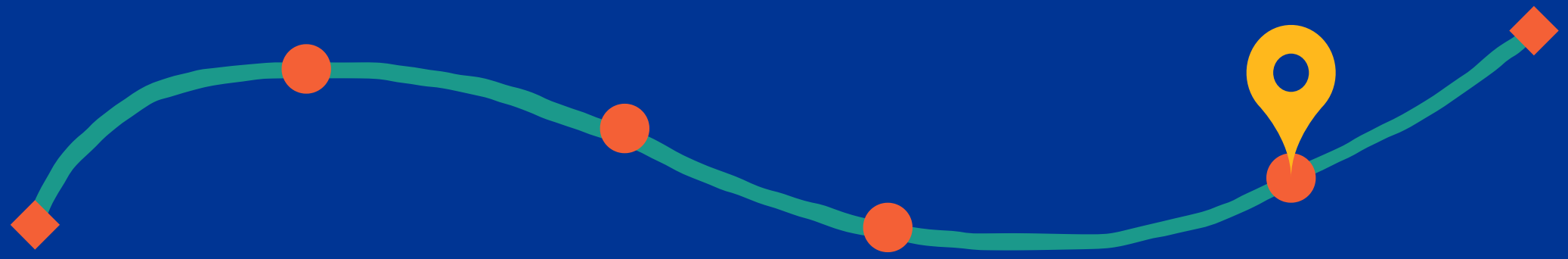
We are here: Active reading

Active reading

Activity time!



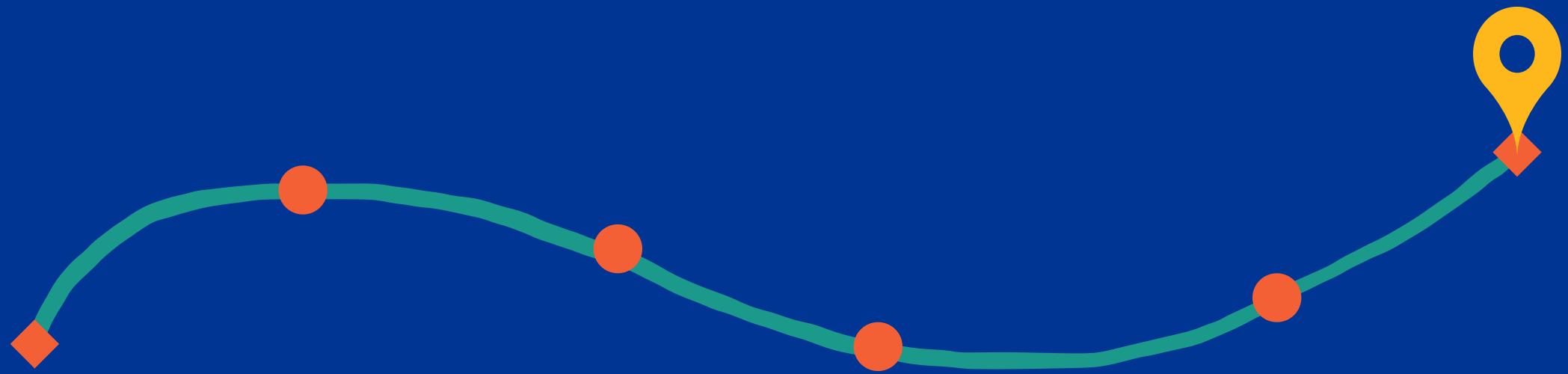
We are here: Break



We are here: Role-based analysis

Role-based analysis

Activity time!



We are here: Closing remarks

Assignment

https://www.aalexmmaldonado.com/biosc1630-2023-fall/assignments/assignment-02.html