Project Proposal

SnaptronUI: Web Based Querying of Exon-Exon Junction Information

Phani Gaddipati
phanigaddipati@gmail.com

Advisor: Ben Langmead

1 ABSTRACT

When a gene is transcribed, introns are spliced out of the original genetic sequence. The splicing of mRNA causes a single stretch of DNA to be expressed in many different gene isoforms. Uncertainty in where the resultant exon-exon junctions are presents difficulty in sequence alignment and genomic studies. Current solutions lack comprehensive information on the source of sample reads. The Intropolis database provides information on ~43 million exon-exon junctions from 21,000 individuals, complete with meta-data. The scale of the dataset makes querying the data extremely difficult. As a result, the information is unavailable to people without a heavy computer science background. Here, a web-app to perform queries on the database is proposed. Allowing common queries, especially "favorite gene" queries, extends the reach of the data to benefit a multitude of studies.

2 INTRODUCTION AND BACKGROUND

Technology advancements in recent years has contributed to an explosive growth in genomic data. Despite the wide availability of large collections, the sheer scale prevents every-day researchers and biologists from using the data. The time and skills required to effectively query the data hinders its use in various studies.

One such collection is the Intropolis database recently produced by the Langmead Lab at Johns Hopkins University. The database provides the results of a study of over 21,000 individual's exon-exon junctions from various tissues.

2.1 EXON-EXON JUNCTIONS

In the process of expressing a protein encoded by DNA, a strand is transcribed into a pre-mRNA strand. Before the mRNA is used by a ribosome to assemble the protein, it undergoes a process of RNA splicing. A subset of introns present in the transcribed sequence are removed by splicesomes, creating exon-exon junctions. [1]

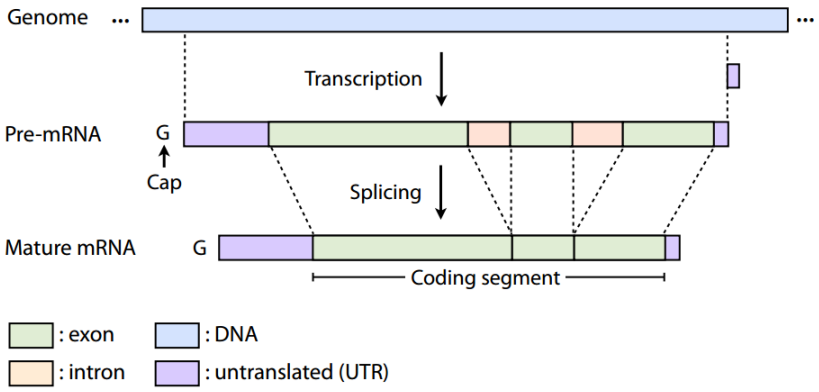


Figure 1: Transcribed mRNA undergoing RNA splicing, introducing exon-exon junctions. Langmead 2015

Differences in RNA splicing results in different gene isoforms, presenting difficulties in gene identification and sequence alignment in genomic studies. A sampled genome contains a permutation of the various genome isoforms – preventing accurate identification without knowledge of the exon-exon junctions.

2.2 Current State

Many current solutions in addressing gene isoforms involves the use of gene annotation databases, such as the Ensembl Gene Set. Derived databases such as Exon-Intron Database (EID) [2] provides information on the presence of exons and introns within genes. The data, created from GenBank, contains almost 250,000 exons over 50,000 genes. However, the locations of the introns is not a part of the database, only the presence in a given gene.

More recently, other exon-exon junction databases have been produced. JuncDB [3], published in 2015, provides exon-exon junction information. The database, targeted towards evolutionary studies, contains 42,347 genes spread over 88 species.

In both of these cases, the source information of the samples are unknown. This limits the applicability in studies where the sample's origin is of importance. If a researcher is interested in liver tissues only, there is no way to filter the database. In the case of JuncDB, web querying is provided, but is restricted to simple Boolean queries.

2.3 Intropolis Database

The Rail-RNA project concurrently processed RNA sequence reads to determine genome positions and identified possible exon-exon junctions. [4] The result is a dataset that provides a summary of ~43 million exon-exon junctions across 21,000 samples. For each junction, the overlaps from various samples are recorded. Associated meta-data for each sample are recorded in a separate table.

The data has potential to be used in many studies. The abundance of meta-data and exact evidence for each junction is unique to the database and would benefit various genomic studies. Exon-exon junction data allows for more accurate splicing-aware sequence read alignments, and aides in the indexing of genes.

The datasets are provided as large flat text files. Unfortunately, the sheer scale presents difficulty in finding relevant information from the database. Usage of the databases requires a heavier computer science background than many people have, and as such, restricts accessibility of the data.

3 PROPOSED WORK

The final product of the work will be a web application to facilitate common queries against the Intropolis database, and if time permits, a gene annotation database. The interface will present the data both in its raw form and visually where appropriate. Further features will be developed based on feedback by interested parties.

The work will focus on "favorite gene queries", catering to many researchers that specialize in a subset of genes. In such queries, the genome interval is already known, and information intersecting with it is of interest. Example questions that the tool would help answer include:

1) Where on a genes' mRNA sequence is a potential exon-exon junction?
2) What evidence is there to support the existence of an exon-exon junction at a particular position?
3) How does the tissue type impact exon-exon junctions?

More specific requirements will be determined by meeting with researchers (bioinformaticians and geneticists) at Johns Hopkins. These will form the basis of an evaluation survey to assess the final product.

3.1 GOALS

3.1.1 Fast Querying

A query should take no longer than a minute, from submission to rendering all showing visuals. Fast querying will allow for more rapid research – more genes can be queried in less time. The query string will be a pair of indexes corresponding to a gene's location in the sequence. Optional parameters should include sample count and coverage.

3.1.2 Meta-Data Based Filtering

The different kinds of meta-data should be enumerated and presented as a filtering mechanism. Such a feature would help refine the queried data to reduce the data to manageable sets. Filters will be pre-enumerated from the complete dataset.

3.1.3 Overlap

When a specific gene location is queried, the resultant location should be visually drawn. An overlay will clearly show where the returned data suggests exon-exon junctions.

3.1.4 (Stretch) Gene Annotation Integration

Use an existing gene annotation such as Ensemble to allow for querying gene's by name. The gene annotation will be used to translate the gene name into the genome location, and the relevant results will be returned.

3.2 EVALUATION

The evaluation will be in the form of a survey. The questions will encompass the following main points in addition to questions specific to the requirements determined. Specifics of the questions will be formulated based on how the final product and previous products are used.

3.2.1 Previous Utilities

- What previous tools, if any, were used to address similar questions?
- How many kinds of queries were accepted (that you used)?
- How does the feature set compare? What is missing and what is beneficial?
- How long did certain queries take to perform with these previous utilities?

3.2.2 SnaptronUI

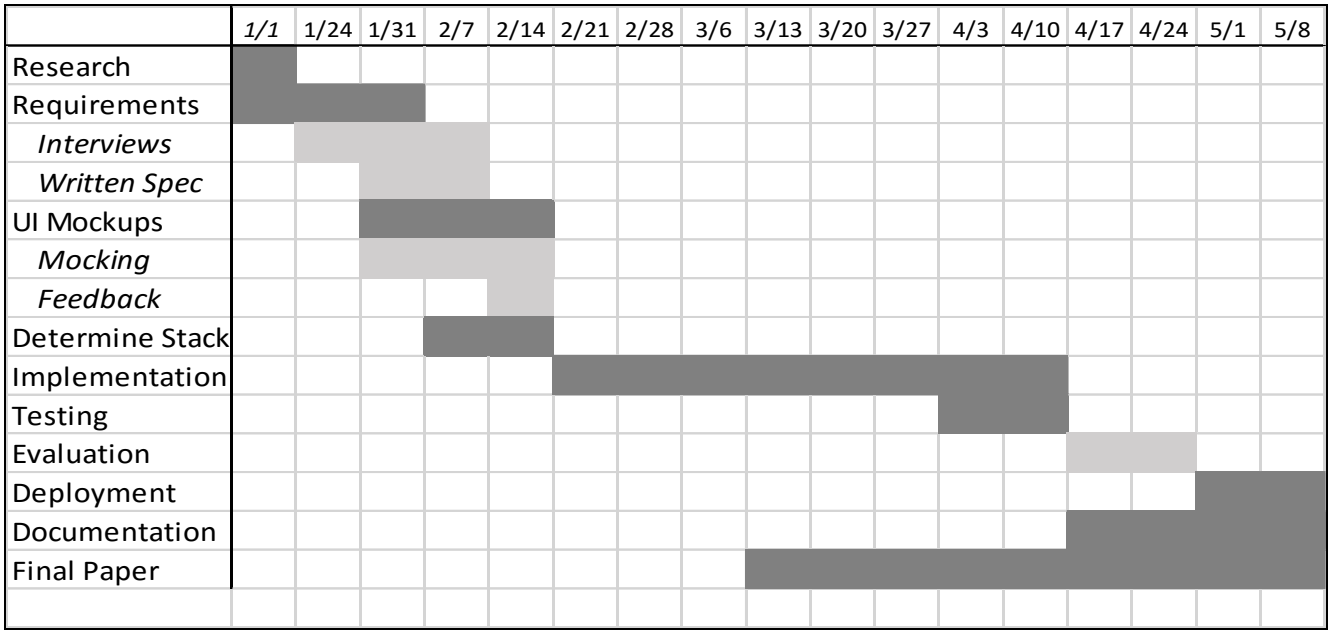
- How long did certain queries take with the new system?
- How many kinds of queries were accepted (that you used)?
- What do you wish was different?
- What is your typical workflow in using this product?

3.2.3 General Impressions

- Is this product preferred over your previous choice of tool?

The survey will be given to the researchers met with in the requirements phase. From the responses, qualitative results will include the overall impression of the utility as well as future development possibilities. Additionally, a *t*-test will be used to determine whether SnaptronUI is better than existing methods holistically and for each query type.

4 TIMELINE



Phase	***Details***
Research	Familiarizing with any existing code and more about the use-cases.
Requirements	Meet with relevant parties to determine how Snaptron can best be useful. This will encompass anything from feature requests to where Snaptron needs to be able to run.
UI Mockups	Draw-up UI sketches to make sure proposed UI will be convenient to use.
Determine Stack	Research and determine which technologies to use to actually implement Snaptron.
Implementation	Core development including server interfacing and UI.
Testing	Testing compatibility, edge cases, and ensuring general robustness for any kind of input.
Evaluation	Assess the product via survey of users determined from the Requirements phase.
Deployment	Streamline the process to deploy the application to some server.
Documentation	Document the code and write a user manual.
Final Paper	Final paper to submit for the project.

5 BIBLIOGRAPHY

[1] S. Clancy, "RNA Splicing: Introns, Exons and Spliceosome," Nature, 2008. [Online]. Available: http://www.nature.com/scitable/topicpage/rna-splicing-introns-exons-and-spliceosome-12375. [Accessed 21 November 2015].

[2] S. Saxonov, I. Daizadeh, A. Fedorov and W. Gilbert, "EID: the Exon–Intron Database—an exhaustive database of protein-coding intron-containing genes," *Nucleic Acids Research,* vol. 28, no. 1, pp. 185-190, 2000.

[3] M. Chorev, L. Guy and L. Carmel, "JuncDB: an exon–exon junction database," *Nucleic Acids Research,* 2015.

[4] A. Nellore, L. Collado-Torres, A. Jaffe, J. Alquicira-Hernández , C. Wilks, J. Pritt, J. Morton, J. Leek and B. Langmead, "Rail-RNA: Scalable analysis of RNA-seq splicing and coverage," 2015.