

Project Proposal: Comparative Genomics

Motivation and rationale:

The Context:

Comparative Genomics involves comparing the genome structures of different species or strains in order to infer their functional properties. Biologists will try to classify the types of structural rearrangements seen between a target and source genome. The types of rearrangements can be used to generate hypotheses about the changes in the function or behaviour of a microorganism that is a result of that genomic rearrangement.

A good example of this is to consider taking a strain of bacteria that is virulent, such as *Staphylococcus aureus* (which is known to cause disease in humans), and a strain that has developed resistance to antibiotics, such as *Methicillin-Resistant Staphylococcus aureus* (MRSA). By looking for inserted segments in the resistant strain when compared to the virulent strains, researchers can identify the sequence segment that is responsible for drug resistance.

The Problem:

Currently visual strategies are used to identify genome rearrangement features. This approach is time consuming and makes it difficult to record the results of a comparison in a computational fashion. It should not come as a surprise that computational approaches are required to classify particular rearrangement features between genomes and to ultimately yield interesting functional predictions. This can subsequently help to prioritise areas for experimental analysis. Consider for example that “in isolation, the human genome [is a] costly but un-interpretable string of three billion or so of A’s, T’s, G’s and C’s.” (Koonin & Galperin, 2003)

One computational approach used in comparative genomics is to perform sequence alignment. Sequence alignment tools produce large textual lists of all the matching areas between the two or more sequences. This makes it difficult for researchers to extract meaningful biological data from the output. Visualization tools are also available that take in this largely incomprehensible data and visually represent all the rearrangement types. However even then a researcher would need to manually identify particular features of interest. Consequently research needs to be conducted in the automatic identification and classification of the types of sequence rearrangements that are of biological interest in sequence comparison data.

Novel approach:

My project aims to explore a novel approach to computationally classifying sequence rearrangement features. I will explore ways of representing the computationally derived comparison data as graphs. Once the comparison data has been represented in graph format, graph theory lends itself well to the discovery of known subgraphs, patterns and motifs. This approach to representing genomic comparison data has not yet been widely explored and so represents a substantial research component of the project.

Aim: To research and develop computational approaches to recognise and classify genome comparisons using graphs based representations.

Objectives:

1. Research methods of representing genome comparison data as graphs.
2. Develop a subgraph recognition algorithms to detect particular rearrangement types.
3. Validate the performance of the algorithms developed in (2) using synthetic rearrangement data.

4. Use the pattern recognition algorithm to detect the rearrangement types in previously characterised biological data and cross-check the results with what is already known.
5. Output a list or populate a database with the results of applying (2) and (3) to bacterial genome comparisons.

Background:

Paper: (Doolittle, 2000), “Uprooting the Tree of Life”

Description: Doolittle’s paper concerns itself with re-discovering the phylogeny (family tree) of microorganisms to account for horizontal gene transfer whose role in the evolution of eukaryotes was overlooked.

Relevance: My data will consist of different strains of bacteria and understanding lateral gene transfer is important for my project as it is largely responsible for the comparison features I will be trying to identify. The paper also provides an interesting biological context to the project.

Paper: (Flanagan, Stevens, Pocock, Lee, & Wipat, 2004), “Ontology for genome comparison and genomic rearrangements.”

Description: This paper briefly looks at the limitations of current comparative sequence analysis tools, and outlines an ontology for describing genomic rearrangements, as well as their biological and evolutionary function.

Relevance: This paper nicely frames the motivation for my project as it outlines the need for automation when it comes to the annotation of biologically meaningful data in comparison sequence data.

Paper: (Pham & Pevzner, 2010) “DRIMM – Syteny: decomposing genomes into evolutionary conserved segments”.

Description: This paper discusses the lack of general purpose computational tools for syteny block identification in highly duplicated genomes. The authors outlined the limitations of the existing syteny block generation algorithms and addressed them by using Prezner’s et al. (2004) A-Bruijn graph approach.

Relevance:

This paper is relevant to my project as it describes a graph approach for identifying rearrangements which is the approach I will be studying. The paper outlines that by using a graph approach the duplications are easier to identify. This further increases the motivation for my project as the paper indicates that using a graph approach works better than other approaches.

Paper: (Flanagan, Pocock, Lee, & Wipat, n.d.) “Logical and Probabilistic Reasoning for Genomic Rearrangement Detection”

Description: This paper describes how comparison data is given to a Bayesian inference network in order to attempt to classify genomic rearrangements.

Relevance: The paper describes one of the approaches that has been taken in the attempt to classify genome rearrangements. I will be using a slightly different approach in my project but our ultimate goal is the same.

Paper: (Sankoff, 2009), “The where and wherefore of evolutionary breakpoints”

Description: This paper explores the idea that breakpoint regions in genomes are more important to chromosomal evolution than the large segments that are conserved between related species.

Relevance: The paper proposes that the location of various breakpoints tend to be in areas of the genome that are responsible for transcription. It also explains how important these rearrangements must be to have survived through various evolutionary pressures. If anything




the paper is relevant because other than the general knowledge it provides it also highlights the importance of identifying and studying gene rearrangements.

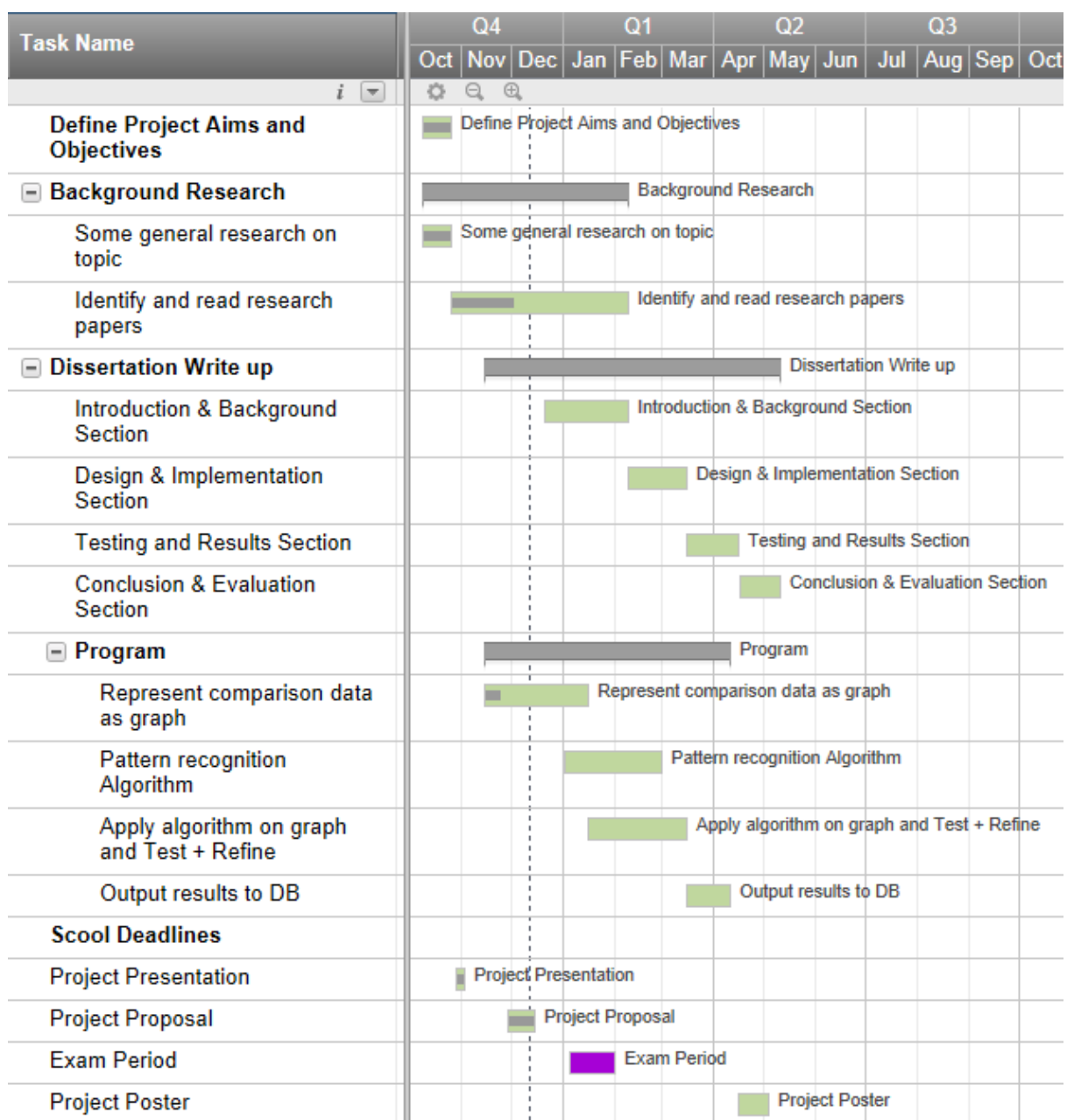
Paper: (L. P. Cordella, n.d.) “An Improved Algorithm for Matching Large Graphs”

Description: This paper describes an improved version of a graph matching algorithm which is used to solve graph isomorphism and subgraph isomorphism problems.

Relevance: The paper addresses an issue that I will face in my project when it comes to finding subgraphs. Taking into account that comparison data will result in a graph that consists of thousands of nodes, finding a low complexity algorithm of low memory requirement will be an important consideration.

Gantt Chart:

KEY	
	Duration
	Complete %
	Not part of Project



Work Plan:

Work done so far:

I started off with some general research around the subject of comparative genomics. This background research involved coming to grips with certain key concepts in genetics such as how Horizontal Gene Transfer works in bacterial genomes. I also spent some time exploring a sequence visualization tool called ACT: Artemis Comparison Tool by the Wellcome Trust Sanger Institute. This program takes in a file which contains the alignment positions of all the matches found between two sequences. The file was generated by a sequence alignment tool called BLAST: Basic Local Alignment Sequence Tool from the National Institutes of Health. I spent some time reading the BLAST manual in order to understand the contents of the file which I will later parse into my program to use as my graph data. At this point it was helpful to start reading various journal papers which addressed the issue of automating the extraction of meaningful biological data from comparison data. I have sketched out on paper what my graph structure could look like for a simple rearrangement type of my choice; I selected a simple instance of an insertion from screenshot from ACT and found the BLAST line responsible for it and parsed it in. My current data subjects are two strains of the bacteria *Staphylococcus aureus*. Bacterial genomes are much smaller than those of eukaryotes and so there is a smaller volume of data to deal with, though it is still very big.

Future work:

My BLAST file contains approximately 8561 lines and I aim to start with a few lines and increase my dataset incrementally. My aim is to research various ways to represent this data as nodes and edges data. The graph could contain several types of edges, and one or more types of nodes, the interaction of which should represent a graph that is capable of being queried. I will do a lot of experimentation with dummy data and throughout the project my graph structure is likely to change as I explore different representations. (See objective one)

As can be seen in my Gantt chart towards February I will move on to my second objective and start to develop a pattern finding algorithm which will identify each rearrangement type. To avoid any noise from the data produced by BLAST, I will create my own synthetic data set containing known examples of each rearrangement type (See objective 3). This can easily be done in accordance to the definitions of each rearrangement type. For example, an insertion can be defined as an area that exists in the target sequence that does not exist in the source sequence. I may also decide to parse in some other file formats that contain additional data such as the genes that are located inside the matching areas. Again the success of the pattern matching algorithm can be measured by using a constrained set of data for which I know the result of. For example I could use a reduced dataset which I know contains four insertions and two deletions. If I swapped around the query and target sequence than my results would have to contain four deletions and two insertions respectively.

A risk associated with using data that increases with complexity and volume is that the algorithm I develop might only be able to handle small data sets. Even if I refine it to perform optimally for a large data set, there is also the risk that my program will only work well with a synthetic data set and doesn't tolerate any noise. In order to accommodate for this I have placed in my Gantt chart an "apply algorithm and refine" phase which coexists with the "algorithm development" phase to illustrate the iterative testing process my program will undergo.

As can be seen in my Gantt chart, the month of January is my exam period and so I have taken this into account when considering the duration of the tasks planned in the months of January.

During the Easter break I have placed my fifth and last objective of outputting the results of the pattern finding algorithm to a database which depending on how far I've gotten I may or may not implement.

References:

- Doolittle, W. F. (2000). Uprooting the tree of life. *Scientific American*, 282(2), 90–5. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10710791>
- Flanagan, K., Pocock, M., Lee, P., & Wipat, A. (n.d.). Logical and Probabilistic Reasoning for Genomic Rearrangement Detection.
- Flanagan, K., Stevens, R., Pocock, M., Lee, P., & Wipat, A. (2004). Ontology for genome comparison and genomic rearrangements. *Comparative and functional genomics*, 5(6-7), 537–44. doi:10.1002/cfg.436
- Koonin, E., & Galperin, M. (2003). *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*. Boston: Kluwer Academic. Retrieved from <http://www.ncbi.nlm.nih.gov/books/NBK20260/>
- L. P. Cordella, P. F. (n.d.). An improved algorithm for matching large graphs. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.101.5342>
- Pham, S. K., & Pevzner, P. a. (2010). DRIMM-Synteny: decomposing genomes into evolutionary conserved segments. *Bioinformatics (Oxford, England)*, 26(20), 2509–16. doi:10.1093/bioinformatics/btq465
- Sankoff, D. (2009). The where and wherefore of evolutionary breakpoints. *Journal of biology*, 8(7), 66. doi:10.1186/jbiol1162