**Bioinformatics Project Proposal: Quantifying Hi-C Data Similarities Between Breast Cancer and Non-Cancer Patients**

**Gabriel Fonseca, Sumitra Lele, Simon Chow**

Histochromatin data is an important class of biological information used to study epigenetics. Epigenetic diseases are a cornerstone of novel biological research. For example, studying different histone chromatin open regions can allow for inferences about cancer. One way to study chromatin data is by looking at correlated sequences in DNA. For example, by looking at sequence similarity, biologists are able to infer which regions are linked, and as such may have roles in chromosomal higher order structures. Using this idea, comparing regions of similarity across different experimental conditions (namely, breast cancer vs healthy cells) can suggest specific regions of DNA implicated in disease (breast cancer).

The paper from the Kingsford lab is an example of a computational method that may be suitable for identifying regions of similarity in the genome (Sauerwald and Kingsford, 2018). In our project, we hope to replicate the research in the paper. To this end, our goal is to assess the efficacy of these methods using different datasets. The idea is to learn whether or not these methods are extendable to different data and, if not, how the methods may be improved or tuned for specific datasets.

The methods in the paper use a dynamic programming approach, similar to an alignment problem. The regions of similarity can be thought of as finding a local alignment across candidate topologically associated (TAD) regions. The score of the similarity regions is calculated using a relative entropy metric. This process is repeated for every pairwise candidate region in the genome. The overall results are reported using statistical models, giving a p-value for each TAD across different experimental conditions. We expect that the runtime for this algorithm might be a hurdle given that the genome is very large. For our analysis, we may only do a subset of them, or employ a heuristic to choose relevant regions to compare.

The dataset we will use will be derived from the Gene Expression Omnibus (which we found has Hi-C data on breast cancer). We may find more datasets later on if our initial analyses are promising. In order to determine how efficient the proposed methods are, we may compare it against existing methods. We hope to find that, if there are specific use cases for different methods, then we should be able to report these in a succinct way. We may, for example, find that the methods used in the paper are more accurate but require higher computational complexity than more typical methods. Although not the goal of our project, if we are able to develop our own methods, we will report these as well.

Our project can be broken down into the following key stages: replicating the algorithms for analyzing the similarity of topology, gathering the data and processing it, and analyzing the results. We believe getting and processing the data should take about one to two weeks, replicating the algorithm would take a little over a month and lastly, the analysis should take about two weeks. Hence, we should be able to complete this project by the end of April.

**References**

1Natalie Sauerwald, Carl Kingsford, Quantifying the similarity of topological domains across normal and cancer human cell types, Bioinformatics, Volume 34, Issue 13, July 2018, Pages i475–i483, <https://doi.org/10.1093/bioinformatics/bty265>

* Gene Expression Omnibus: <https://www.ncbi.nlm.nih.gov/geo/>