# ECS253 - Homework 2a

Luiz Carlos Irber Jr
*UC Davis*
lcirberjr@ucdavis.edu

## PROJECT OVERVIEW

A genome can be represented computationally as a sequence of letters representing nucleotide bases, a string over the alphabet {A, C, G, T}. Current technology can't sequence an entire genome at once, but can sequence smaller randomly sampled subsets (called reads), usually ranging from 100 to 10k bases, with varying degrees of error (1-15%). The computational challenge is the fragment assembly problem: given a multiset of reads from a genome, reconstruct the original genome. Complexity arises from regions with repeated patterns, sequencing errors, and multiple chromosomes. Methods to solve this problems typically use assembly graphs, where reads are connected when there is an overlap between them; the original sequence is estimated by walks through this graph.

Prior work on the problem has assumed the availability of the complete assembly graph; that is, the algorithms are *offline*. While this approach has been sufficient at the scale of sequencing projects thus far, the growing volume of sequence data suggests that an *online* approach will eventually need to be adopted. To that end, we propose to develop an arrival model of the assembly graph parameterized by depth of coverage, sequencing error rate, and an estimator of the underlying genomic complexity. Such a model would aid in the development of new online assembly strategies, providing the theoretical groundwork for a new class of assemblers.

Since this is an initial step for many biological experiments, the quality of results is essential to any posterior analysis, be it a microbiome study or how a new drug interacts with antibodies during treatment. Metagenomics assembly has particular opportunity for discovery in this area; in these studies, a diverse community of potentially hundreds of thousands of microorganisms are sequenced, with varying levels of genome similarity. The assembly graph that results is extremely complex, and existing assemblers do not utilize strategies capable of coping with the dozens of terabytes of data. A better model can lead to improved methods, especially when facing the increasing complexity and amount of data being generated.

## LITERATURE REVIEW

*Scaling metagenome sequence assembly with probabilistic de Bruijn graphs*

Pell et al. (2012) describes a probabilistic de Bruijn graph representation for sequencing reads using Bloom filters, allowing memory-efficient storage. Percolation theory is used to study how false positive rates can lead to long-range connectivity: below the percolation threshold only local elaborations of the graph structure occur. The de Bruijn graph representation works well with site percolation, since the maximum number of k-mers is fixed ($4^k$ for an *ACGT*) alphabet) and k-mer occurrence defines the edge connectivity. In the context of metagenomes, the article

also discusses graph partitioning using this representation, since components can be assembled separately and with less resources than the full graph.

For our project the site percolation discussion is interesting: can we use percolation as a measure of sequencing saturation? This depends on what our input data looks like (genomes and metagenomes have different properties, for example), but it might work in some contexts.

*Anomaly Detection in Streaming Sensor Data*

Pawling et al. (2008) describes WIPER, a proof-of-concept decision support system for emergency response managers. It uses cell phone network data to detect anomalies using a streaming approach, thus treating cell phones as sensors. WIPER is a dynamic data driven application system (DDDAS), meaning it is a framework in which running simulations incorporate data from a sensor network to improve accuracy. The article focus on the detection and alert system, reviewing previous work on outlier detection and clustering, with special attention to streaming algorithms, which typically use no more than $O(\log n)$, where $n$ is the number of data items. It also explores percolation theory for spatial anomaly discovery. Finally, three data mining techniques have been implemented: 1) a model using a Markov modulated Poisson process technique, 2) a method for spatial analysis based on percolation theory, and 3) a method for spatial analysis using online hybrid clustering.

In the genomic sequencing context we can considers sequencers as sensors, and the techniques described in this article seems to be applicable in this context. Sequencing errors can be treated as anomalies, with the caveat that we might not know what is the underlying distribution for the input (or have preexisting sequencing data from similar organisms to train the normal case).

*Generating Application-Specific Benchmark Models for Complex Systems*

Wang and Provan (2008) describes an automated generator for benchmark models. It uses a domain-analysis algorithm to build a system topology that best matches the real world system topology. The topology is generated both from explanatory and descriptive models, where the first one depend on a set of candidate algorithms with parameters optimized iteratively until some of them match the real-world network properties, and the second one depends on an integer parameter $d$, increased until the generated graph matches the properties of the real-word graph with sufficient fidelity. The descriptive model also has higher computational cost than exploratory approaches.

The takeaway of this article is more on the methodological side, on how we can set the model validation and generation for our model. Given the availability of already assembled genomes (and the underlying reads used for assembly), we can use them to analyze key properties and see if our model generates similar results. It is still focused on exponential and power law degree distributions, which don't fit so well with our project, and so we can introduce our model to compare with the other candidate algorithms to see if it performs better.

## REFERENCES

Pawling, Alec, Ping Yan, Julián Candia, Tim Schoenharl, and Greg Madey. 2008. "Anomaly Detection in Streaming Sensor Data." *Intelligent Techniques for Warehousing and Mining Sensor Network Data*: 99–117.

Pell, Jason, Arend Hintze, Rosangela Canino-Koning, Adina Howe, James M. Tiedje, and C. Titus Brown. 2012. "Scaling Metagenome Sequence Assembly with Probabilistic de Bruijn Graphs." *Proceedings of the National Academy of Sciences* 109(33): 13272–7.

Wang, Jun, and Gregory M. Provan. 2008. "Generating Application-Specific Benchmark Models for Complex Systems." In *AAAI*, 566–71.