Growing Assembly Graphs: A Brief Literature Review

Camille Scott

May 5, 2016

Abstract

Significant prior work has been done on the statistics of the sequencing process itself, where the sequencing requirements for sampling a complete genome are explored; the construction of assembly graphs from sampled fragments; and the assembly process itself, which is mostly an exercise in the interrogation of these graphs. To better understand our problem, I will review several papers that span these topics. The first is one of the original papers describing the sequencing coverage statistics; the second, a detailed treatment of the assembly problem through the lens of assembly graphs' potentials to fully represent the information in a genome; and the third, an examination of percolation theory in probabilistic de Bruijn graphs which serves as a useful jumping-off point for our project.

1 Original Project Proposal

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 cant 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.

2 Sequencing Statistics

Paper: J. C. Roach, "Random subcloning.," *Genome Research*, vol. 5, pp. 464–473, Dec. 1995

Roach's work extended a coverage model originally proposed by Lander and Waterman [2]; the prior model is accurate at low redundancy, but not at the higher redundancies present in shotgun studies. This paper came from the days of "strategic sequencing," when sequencing technology was still in its infancy and optimizing cost versus information was extremely important for genomics projects. While we no longer have this concern due to the proliferation of extremely cheap shotgun sequencing, Roach's coverage equations are still important for understanding how genomic information increases as additional fragments are added, and as such can help us understand how assembly graphs grow as more fragments are added.

This model starts by assuming a linear discrete genome G, from which n fragments of length L are generated. Fragments cannot overlap the end of G, so there are $G_e = G - L + 1$ available start sites for fragments. Redundancy R is $\frac{nL}{G}$. We assume that T bases of a pair of fragments are necessary to detect an overlap. They then look at the domain space of spacings between start sites D_k. They show that the density function for these lengths is:

$$f_{D_k}(x) = n(1 - \frac{1}{G_e})^{n-1}$$

From this, and the observation that a gap occurs at a site S_k iff $D_k > L - T$, they go on to calculate the probability of a gap following a fragment, where $f_G = fracL - TG_e$ is the effective fractional target coverage per fragment (in other words, the contribution of each fragment), as:

$$P_{gap} \approx \int_{L-T}^{G} f_{D_k}(x)dx = (1 - f_G)^n$$

From P_{gap}, they then go on to derive: the expected number of gaps $E(N_{gaps})$, the expected number of islands $E(N_{islands})$, the expected fragments per island, and several models of the distribution of lengths of the islands.

While this may seem somewhat obtuse at first glance, its importance is due it being directly translatable to a number of assembly graph models by transforming P_{gap} to the probability of edge arrival. For example, existing de Bruijn graph models set L to be equal to the k-mer size $K+1$ [3]; the analyses follow directly. The models provided by Lander, Waterman, and Roach will then directly inform any graph model we build.

3 Assembly Feasibility

Paper: G. Bresler, M. Bresler, and D. Tse, "Optimal assembly for high throughput shotgun sequencing," *BMC Bioinformatics*, vol. 14, no. 5, pp. 1–13, 2013

This work provides a new take on Roach's model. Roach's model does not consider repeats in the DNA sequence, which add considerable complexity to assembly graphs and the assembly process. In the absence of repeats, assembly graphs would be collections of disjoint components which nodes of degree 0, 1, or 2. Repeats introduce cycles, and more directly, ambiguity to the ordering of traversals which determine the assembled sequence. Bresler *et al.* study the **feasibility** of an assembly, whether it is possible to construct the original sequence from a set of reads, and the **optimality** of an algorithm in terms of their ability to reconstruct the original sequence when it is theoretically feasible. Feasibility is a "measure of the intrinsic information each read provides about the DNA sequence," and is closely related to the length and frequency of repeats in the original genome. This work features a number of plots of the performance of specific graph and assembly paradigms versus the lower bound on feasibility for increasing L.

The introduction of repetitive sequence is another step toward our growth model. Assembly graphs will not be modeled exactly as traditional random graphs, because there is some specific underlying sequence which they aim to represent; the properties of that sequence, such as its repeat content, will need to form the parameters of out model. It's also worth noting that the authors identify a two critical points for increasing fragment length at which feasibility is achieved with reasonable coverage depths; we may be able to find some relationship between these critical points and the geometric phase transition observed in [4] and described in the next section.

4 Percolation in de Bruijn Graphs

Paper: J. Pell, A. Hintze, R. Canino-Koning, A. Howe, J. M. Tiedje, and C. T. Brown, "Scaling metagenome sequence assembly with probabilistic de Bruijn graphs," *Proceedings of the National Academy of Sciences*, vol. 109, no. 33, pp. 13272–13277, 2012

This work was primarily concerned with the employing a probabilistic data structure for the representation of a de Bruijn graph. Previous de Bruijn graph implementations stored k-mers exactly, requiring massive memory. Here, the authors use a Bloom filter [5] to only store the hashed k-mers; traversal is accomplished by exhaustively checking for the presence of neighbors. Because each node can have at most eight neighbors (four in either direction), the additional cost is minimal. This advancement radically reduced the memory required to store a de Bruijn graph, while introducing a new challenge to contend with: Bloom filters have a false positive rate, parameterized by the size of their underlying table and the number of elements they contain. In a network sense, this means that some nodes will gain neighbors corresponding to sequences which do not exist in the original set of fragments.

The practical goal of this work was to divide an extremely large set of fragments into approachable chunks. In metagenomics, the set of fragments contains genomes from potentially millions of species; in a ”perfect” data set, where genome share no sequence content and there is no error, each of these genomes would be represented by its own connected component in the assembly graph. An effective solution to decomposing the metagenome assembly problem is then to find all the disconnected components in the assembly graph and divide up the fragments based on the component to which they correspond. To use the probabilistic de Bruijn graph for this task, it was important to determine whether false positives in the data structure could cause otherwise disjoint components to become erroneously connected. It was observed that for increasing false positive rates, the effect on the connectivity was similar to a geometric phase transition: at false positives rates of about 0.18, the average component size begins to grow rapidly. This is first shown via simulation, and then studied with percolation theory. They used a site percolation model where the probability p of a site being active is equivalent to the false positive rate of the bloom filter.

This site percolation model should be relatively close to our model. However, rather than taking the false positive rate of the Bloom filter as p, we will need to leverage Roach's model for fragment coverage. Intuitively, we expect to see similar behavior with respect to component size with increasing sequencing error as Pell *et al.* did with respect to increasing error in the Bloom filter. Between this and Bresler *et al.*'s critical points for assembly feasibility, we hope to identify other important phase transitions in the assembly graph.

References

[1] J. C. Roach, “Random subcloning.,” *Genome Research*, vol. 5, pp. 464–473, Dec. 1995.

[2] E. S. Lander and M. S. Waterman, “Genomic mapping by fingerprinting random clones: a mathematical analysis,” *Genomics*, vol. 2, pp. 231–239, Apr. 1988.

[3] G. Bresler, M. Bresler, and D. Tse, “Optimal assembly for high throughput shotgun sequencing,” *BMC Bioinformatics*, vol. 14, no. 5, pp. 1–13, 2013.

[4] J. Pell, A. Hintze, R. Canino-Koning, A. Howe, J. M. Tiedje, and C. T. Brown, “Scaling metagenome sequence assembly with probabilistic de Bruijn graphs,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 33, pp. 13272–13277, 2012.

[5] A. Broder and M. Mitzenmacher, “Network applications of bloom filters: A survey,” *Internet mathematics*, vol. 1, no. 4, pp. 485–509, 2004.