A Graph Theoretical Approach to DNA Fragment Assembly

Jonathan Kaptcianos[‡]
Saint Michael's College
Colchester, Vermont 05439 USA

Received: September 19, 2007 *Accepted: April 1, 2008*

ABSTRACT

Built on prior work on sequencing by hybridization, fragment assembly is a newly explored method of determining whether or not a reassembled strand of DNA matches the original strand. One particular way to analyze this method is by using concepts from graph theory. By constructing data models based on these ideas, it is possible to come to various conclusions about the original problem regarding reassembled strands of DNA. In this paper we will detail this approach to DNA fragment assembly and present some related graph theoretical proofs in the process, including the BEST theorem. Further, we will explore the Eulerian superpath problem and its role in aiding fragment assembly, in addition to other recent applications of graph theory in the field of bioinformatics.

I. Introduction

In recent years, scientists and researchers have emphasized DNA sequencing and fragment assembly with the hopes of enhancing their abilities to reconstruct full strands of DNA based on the pieces of data they are able to record. Complications arise with fragment assembly due to imperfect data sets. Strands are often riddled with repeats and come in varying sizes. As a result, configuring the image of the original genome is not as easy as fitting one puzzle piece into the next.

We will discuss a recently discovered approach to DNA fragment assembly that uses components from graph theory, including de Bruijn graphs and Eulerian circuits, to successfully compensate for some of these errors. Specifically, this approach, originating from work done with sequencing by hybridization, consists of constructing various directed graphs based on provided DNA data, and counting possible Eulerian circuits in these graphs.

This paper will provide some preliminary background information from graph theory. In addition to several definitions, we will prove two theorems, the BEST theorem and the matrix tree theorem, both of which are used in counting Eulerian circuits in a directed graph.

Further, this paper will briefly survey the current state of DNA sequencing and fragment assembly, such as the problems that arise, and how they are addressed using graph theory. We will see how graphs are being constructed using DNA data, and how these graphs inform the problem. Our primary emphasis falls on the Eulerian superpath problem of Pevzner, Tang and Waterman in [1-2], which is a more recent attempt to simplify DNA graphs through a series of transformations on the graph. Such transformations include different edge detachments for cases of single and multiple edges in a directed Eulerian graph.

The concluding section discusses the decomposition of DNA sequencing with nanopores, as detailed in Bokhari and Sauer [3], and the role that graph theory plays in resolving problems that arise. This is a more recent application of graph theory being put to use in the field of bioinformatics.

II. Graph Theory Definitions

Graph theory is a field of mathematics which has many applications in the world of science. We assume the reader is familiar with the basic foundations of graph theory, such as that found in Tucker

[‡] jkaptcianos@smcvt.edu

[4], and West [5]. The basic terminology here follows that of West [5].

One can move through the elements of a graph in several different ways. A walk is any journey through a graph which produces a list of vertices and edges such that an edge connects the vertices on either side of it. A trail is a walk where no edges are repeated. A closed trail is a circuit, and the list of vertices is expressed in cyclic order. Further, a path is a trail in which no vertices are repeated such that the vertices can be listed in consecutive order as one is adjacent to the next. In the case of a directed graph, the edges of a trail or circuit must be consistently oriented.

A trail or circuit which traces every edge in a graph once and exactly once is considered Eulerian. Eulerian trails and circuits are allowed to pass through vertices in a graph more than once. The number of Eulerian circuits in a graph can be computed with a formula generated from the BEST theorem, which will be proven later in this paper.

A final definition from the field of graph theory which we use in this paper is that of de Bruijn graphs. A de Bruijn graph is a directed graph with vertices that represents sequences of symbols from an alphabet, and edges that indicate where the sequence may overlap, as defined in de Bruijn [6].

The construction of this graph is dependent on the set of *l*-length pieces, or fragments, from the particular sequence at hand. Each vertex is labeled by a fragment of length *l* – 1 and the directed edge existing between two vertices represents one of the *l*-length fragments. Specifically, the first symbol in the fragment comes from the vertex that sends the edge, and similarly, the last symbol comes from the receiving vertex. Thus, the remaining symbols that both vertices contain are found labeling the edges.

For example, we construct a de Bruijn graph for the sequence “0110101” using fragments of length *l* = 3. The four triples that are present in the sequence are 011, 110, 101 (which appears twice), and 010. Figure 1 shows the de Bruijn graph that coincides with this sequence. Notice the vertices represent the first and last subsequence of length 2 in each triple, and the directed edge between two vertices represent the respective triple.

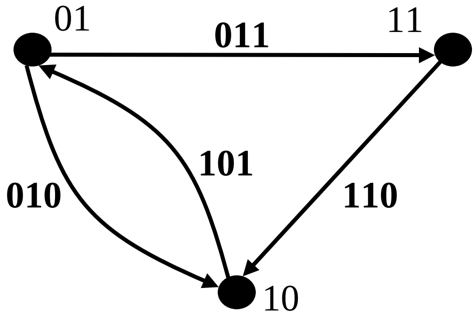

Figure 1. A de Bruijn Graph for the sequence “0110101” with fragments of length 4.

III. DNA Sequencing and Fragment Assembly

For simplicity, consider DNA sequencing to be much like the process of constructing a children’s puzzle, as expressed by Pevzner, Tang, and Waterman [1]. The only difference lies in the fact that scientists are dealing with hundreds of pieces of varying sizes, and some pieces are exactly identical to one another. The “puzzle pieces” are various strands of DNA reads and the completed puzzle is what the entire original strand, or genome, looks like. The provided information on this aspect of Bioinformatics comes from various sections of the text by Jones and Pevzner [7].

Specifically, DNA fragment assembly is a lab technique which determines the configuration of an entire genome given a series of viewable reads from that particular genome. Since a full strand of DNA consists of millions of nucleotides, there is no way to visibly capture the structure in its entirety. With today’s technology, scientists and researchers are able to look at fragments of DNA anywhere from 500 to 1200 nucleotides long. These fragments all consist of the assortment of the four types of bases which make up DNA: adenine (A), thymine (T), guanine (G), and cytosine (C). There could be multiple ways to reconstruct

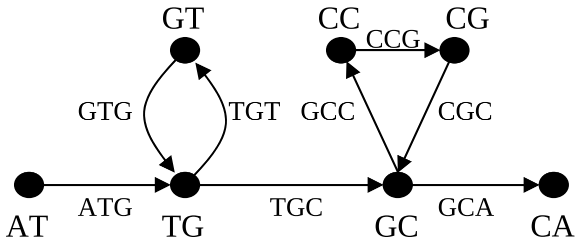

Figure 2. A de Bruijn graph for the sequence ATGTGCCGCA.

the original strand out of the fragment pieces but only one of them is correct.

There are many problems and complications that arise in fragment assembly which make it much more difficult than simply constructing a puzzle. Within the portrayal of the viewable reads there exists a rate in which errors are produced by sequencing machines. Anywhere from 1% to 3% of DNA reads resulting from sequencing may contain errors and do not appropriately represent a part of the full strand.

Further, another problem lies within the fact that DNA is a double-stranded structure. Based on the work of Watson and Crick [8], we know that with each single strand of DNA there exists a complement, as A matches with T, and C is complementary with G. Thus, upon looking at reads from a strand of DNA, we are unable to determine whether or not a fragment is coming from the desired strand, or its complement.

While the error rates and complement confusions do hinder some steps in DNA sequencing and fragment assembly, the major area of concern and confusion comes from repeated sequences. The human genome has a large number of sequences that repeat multiple times. Further, if a repeating sequence is larger than the size of the viewable reads, it would make construction of the genome almost impossible. For example, consider a particular genome which has a repeating sequence spanning a stretch of 2000 nucleotides. No sequencing machines today can view a read of this length, so there is no way to ever see the repeat in its entirety.

Most attempts at correcting errors in fragment assembly aim to attack the problem of repeats. Previous approaches for doing so followed the “overlap-layout-consensus” algorithm. The first step involves matching all possible reads and finding any overlapping. This is done by looking at the beginning sequence of one read and the ending of another. The layout step is the construction of the strand, and it proves to be the most difficult. Keeping repeats in mind, an attempt is made to find the order of reads along a full strand of DNA and put them together. The last component of this algorithm is deriving how the sequence will appear based on the layout produced in the previous step. The approach we will discuss in this paper abandons the traditional outline of this three step process, and in doing so, becomes a matter of probability.

IV. Resembling Strands of DNA with de Bruijn Graphs

Sequencing by hybridization is a lab technique which similarly looks at fragments of DNA, known as DNA arrays, but is not to be confused with fragment assembly, as they are different. Specifically, sequencing by hybridization (SBH) relies mostly on the binding of an unknown target fragment of DNA with an array of shorter synthetic fragments known as probes. These probes are anywhere between 8 and 30 nucleotides in length, and they bind to the target based on the Watson-Crick complements, as noted in Pevzner [9].

However, it was a graph theory approach to SBH by Idury and Waterman [10] that introduced similar approaches in the field of DNA fragment assembly. In their article, the authors defined the rules to construct a directed graph based on a sequence of DNA. More specifically, these rules were for constructing a de Bruijn graph based on DNA pieces of k-length. The graph had vertices labeled by fragment of length *k*-1, and edges labeled by fragments of length *k*, which connected two adjacent vertices in the sequence.

Figure 2 shows a de Bruijn graph for the simple sequence ATGTGCCGCA, as it was used by Idury and Waterman [10]. Here, the desired fragments of DNA are of length 3. Thus, the vertices are labeled by fragments of length 2 and edges labeled by fragments of length 3, representing the particular triple from the sequence. Notice how there is only one possible Eulerian trail in the graph above, as this is just a simple case. Various work done by Pevzner in [9, 11] has developed these ideas further.

Now, we will construct another de Bruijn graph using the same rules as above. Given the set of fragments *S* = {ATG, TGG, TGC, GTG, GGC, GCA, GCG, CGT}, the graph would similarly possess vertices of length 2 and edges of length 3. However, the graph produced for this set of reads is more complex than the previous graph. Specifically, there are two different closed trails through the entire graph, producing the sequences ATGGCGTGCA and ATGCGTGGCA. See Figure 3, and note that the extra dotted edge is added to connect the first and last vertices AT and CA, creating an Eulerian digraph.

As previously mentioned, this approach to DNA fragment assembly, referred to by some as the Eulerian path approach, creates a question of probability. With this example above, we see that this given set of fragments constructs two different sequences, based on there being two differing Eulerian circuits. Therefore, there is a ½ possibility that we will resemble the actual genome, depending on which Eulerian circuit we follow. Since we added the extra edge making the graph closed, we are able to count Eulerian circuits as opposed to trails.

More generally, when we are given complete and errorless data, we find that the probability of reconstructing the original strand of DNA given a set of fragments is:

$$\frac{1}{\textit{total \# of Eulerian circuits in the de Bruijn graph}}$$

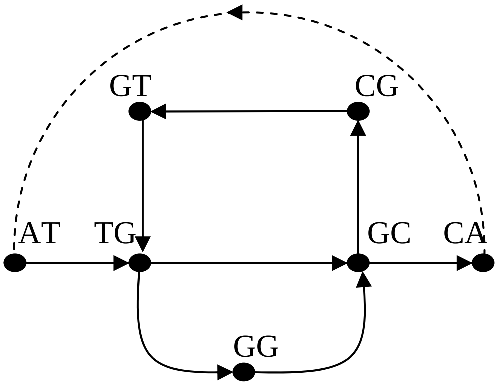

Figure 3. The Graph for the set of reads **S**.

With that, there is a theorem used to count Eulerian circuits in a directed graph, known as the BEST theorem, which we will prove, followed by another related theorem and proof.

V. The BEST Theorem

There exists a formula for computation of Eulerian circuits in a directed graph. Named after its inventors, de Bruijn, van Aardenne-Ehrenfest, Smith, and Tutte, [12-13], the BEST Theorem reads as follows:

Theorem 5.1. Given a connected directed graph ***G*** and set of vertices *V*(***G***) = $\{\nu_1,\dots,\nu_n\}$ all of even degree, the number of Eulerian circuits |*s*(***G***)| is expressed as the following, where |t_i(***G***)| is the number of spanning trees rooted towards any vertex ν_i in ***G***:

$$\left|s(G)\right| = \left|t_i\left(G\right)\right| \prod\nolimits_{j=1}^{n} (d^+(v_j) - 1)!\,.$$

The following proof for this theorem is indicated in Bollobás [14].

Proof: We are given a directed multigraph ***G*** with a vertex set *V*(***G***) = $\{\nu_1,\dots,\nu_n\}$ and the outdegree and indegree of a vertex ν_i being equivalent, which is denoted $d^+(\nu_i) = d^-(\nu_i)$. If these conditions are met, then we know that ***G*** has at least one directed Euler circuit, by a theorem found in Bollobás [14].

Further, let's assign E as the set of directed Euler circuits, and E_i as the set of