Threading Through the Breast Cancer Genome with PacBio Sequencing Data

Marley Alford[1,2], Maria Nattestad[1], Michael C. Schatz[1]
[1] Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11743
[2] Department of Mathematics, Bard College, Annandale-On-Hudson, NY 12504

Introduction

Structural variations in the genome account for some of the most deadly features associated with aggressive cancers such as the HER2+ (amplified) breast cancers (Navin, et al., 2011). Structural variations are genetic rearrangements involving DNA fragments at least 50 base pairs long. They include insertion, deletion, duplication, inversion, inter-, and intrachromosomal translocations (Baker, 2012). Some HER2+ breast cancers are characterized by massive duplication in the HER2 region of chromosome 8, with as many as 20 extra copies of a gene. Twenty percent of breast cancers are HER2 amplified. Many studies on HER2 amplification use SK-BR-3, a model cell line from a patient with this often-fatal cancer. Though HER2+ breast cancer is already known to contain multiple amplifications and deletions, a new study from Dr. Michael Schatz's research group at Cold Spring Harbor Laboratory intended to learn more, mapping all structural variations in order to retrace the steps that prompted their creation. The analysis involves several processes: first the DNA is sequenced using PacBio long-read sequencing; then the structural break-points are ascertained using "split-read" alignments; from that information, structural variations are identified and analyzed; and finally, we have developed a new "genome threading" algorithm to retrace the path each fragment of DNA took in order to create those structural variations. Similar projects have been using Illumina short-read sequencing for the first step because of its lower error rate and lower costs. The PacBio long-read method is more accurate, however, especially for sequencing cancer genomes (Roberts, Carneiro, and Schatz, 2013). This is because long reads (which average 10 Kb, though some approach 100Kb) have a greater chance of spanning long repetitive sections of DNA, compared to 200 bp Illumina reads, which could easily map to the wrong places, erroneously showing large copy numbers in one read of a repeating sequence (Chaisson, et al., 2014). The final step, using DNA threading to retrace the history of the genome, is a step that has never before been successfully attempted. Normally, the most information researchers can attain from DNA sequencing is copy numbers and breakpoints (Baker, 2012). The Schatz lab is now using this information to reach a step further, reconstructing the history of how the original genome attained its new, mutated form. My project this summer was to create a computer algorithm to perform this final step: mapping out the possible DNA fragment paths using breakpoint and copy number data. This new step will have major implications for the breast cancer and cancer fields, as well as any DNA sequencing research in general.

Methods

The DNA threading algorithm is calculated and visualized using graphs where nodes represent DNA fragments, and edges show the connections between these fragments. To maintain the orientation of the DNA sequences, the graph implementation represents DNA fragments as nodes with “start” and “end” ports: traversing a node from “start” to “end” would be the original, forward strand of the DNA, while traversing from “end” to “start” would represent an inversion. The nodes are connected by weighted edges from the port of one node to the port of another. These edges connect the nodes in order to convey the path of DNA fragment mapping, while their weights represent the copy numbers. Each graph starts and ends at the “Portal” node. The “Portal” node is an abstraction to represent that our current graph is only a small part of the larger genome, and that there may be more DNA before and after the DNA sequence of the graph. It is necessary to have this node to perform calculations on the graph, even though it does not exist in the real DNA.

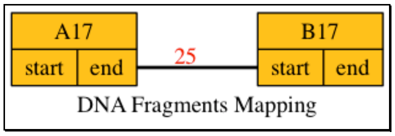


fig 1. An example of a simple graph. A17 and B17 are nodes on chromosome 17. The 25 is the edge-weight, representing how many reads span from sequence A17 to B17.

The Algorithm:

The graph interface chooses the most likely history of DNA mapping using three important tasks: the first is to find all the paths in the graph from Portal to Portal using a depth-first-search of the graph. Once discovered, the algorithm searches for the path with the longest sequence preserved from the original genome. The third task is to subtract this path from the list of all paths by subtracting its minimum capacity edge-weight. The 2nd and 3rd steps are repeated until no paths are left. The path containing the “longest uninterrupted” sequence is that which most closely resembles the original genome sequence. The subtraction of a path is achieved by subtracting its minimum-weighted edge from all edges in the path (Maurice, 1960). This eliminates the path since the minimum edge has been reduced to a value of 0. Then the algorithm finds new paths from the remaining available. With each subtraction, the algorithm is forced to find a path with a slightly smaller original sequence, meaning that structural variation has occurred and created this next path. Thus, by finding each longest path and subtracting it, the algorithm gathers a list of structural variations to the genome in the order of occurrence. Also important is the task of visualizing these possible mapping routes. The program does this by making a graph with each maximum capacity edge represented by a different color (see **fig 5**).

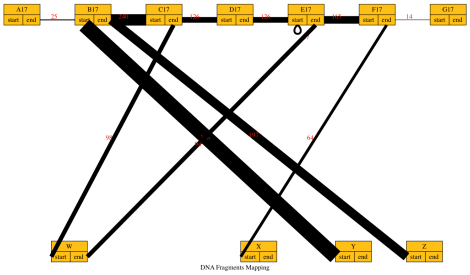
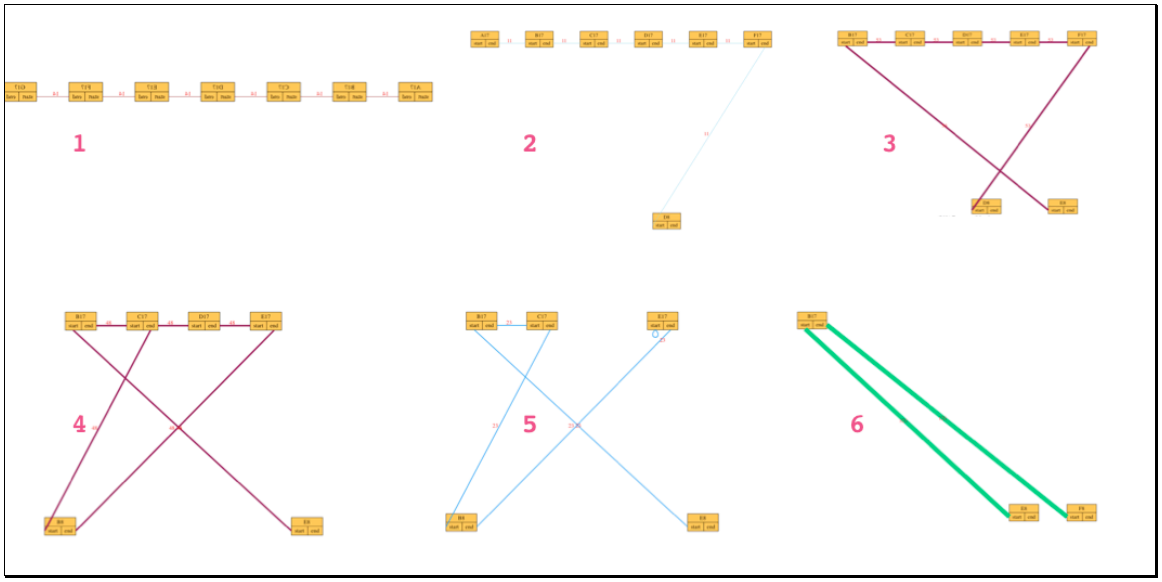


fig 2. (left) Graph of the Her2 region. The nodes in the top row are all from chromosome 17, while those in the bottom are from chromosome 8. The interchromosomal translocations are visible stretching from the top row to the bottom. (next page) With each iteration the algorithm picks paths with smaller and smaller sequences preserved from chromosome 17. Notice, in 1 the whole sequence is preserved, while in 6 only one node from chromosome 17 is left.



Results

Analysis of algorithm efficiency:

If the number of nodes in a graph is N, the efficiency of finding the maximum capacity edges is ***a*** in the worst case (refer to **fig 3**). This is because that process is comprised of three sub-processes, 1) finding all paths in a graph, 2) the subtraction of longest uninterrupted paths (**fig. 2**), and 3) repeating those two processes until all paths are subtracted.

a. $O(2^N)$

b. $4N \Rightarrow N$

c. $O(N2^N) = 2^N + 2^{N-1} + 2^{N-2} + 2^{N-3} + \cdots + 2^{N-N} = 2^{N+1} - 1 \Rightarrow 2^N$

fig 3.

In a complete graph, where all nodes are connected to each other, the task of finding all paths would have an efficiency of ***a***. While our data has only an average of 3 edge connections per node, with 4 being the maximum that is biologically possible, we still use the efficiency of a complete graph as our worst-case estimate. The second task, the minimum weighted edge subtraction, requires finding all remaining paths again after each subtraction; thus the subtraction is performed as many times as there are viable paths to subtract from. As opposed to a complete graph where the number of edges is O(N^2), our model necessarily restrains each node to a maximum of 4 edges, making the total number of edges O(N), ***b***. Therefore, we multiply the efficiency of finding all paths (***a***) to the number of edges (***b***) to obtain the overall efficiency of the first two tasks, ***c***. This completes the algorithm. Currently the algorithm has been used to map only the HER2 region, which has 13 nodes, taking an average of 0.2-0.6 seconds to complete. While the whole genome data merely has 318 nodes, it may be necessary to break up the data for maximum speed of analysis.

Analysis of HER2 region:

The algorithm sequencing and analysis functions determined translocations and sequence coverage at the breakpoints, allowing for the automatic population of the graph with edge weights directly from the data. The node information includes ID numbers for each node (the position in the genome), edge information lists breakpoints, coverage, and other details to identify the particular structural variation for each. Every breakpoint in the edge file creates three new edges between two nodes: two are connections within a node, and the third is a new edge connecting from one node to the next. When a new breakpoint is formed, there is information saying which ID numbers have been connected, thus we are able to see which two nodes have been joined. According to the “strand” data of the edge, we can then see which ports of each node are involved in the connection, and thus the direction in which the DNA will be read. Finally the “spans” give the coverage, or edge weights of edges adjacent to breakpoints. When the algorithm reads in this data, it is used to make a graph (**fig 4**).

chromosome1	ID	chromosome2	ID	strand1	strand2	split	span1	span2
8	1584290	17	1652015	+	-	20	12	6

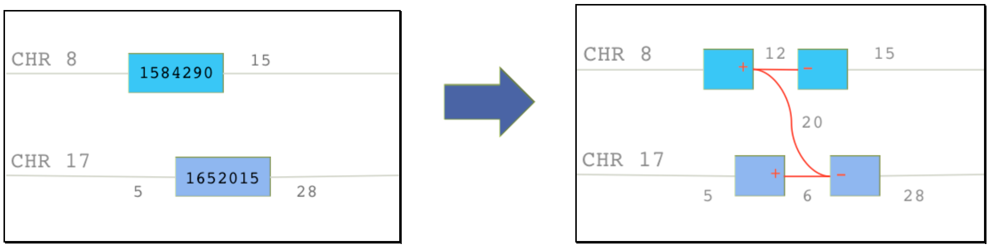


fig. 4 Using the data above, the algorithm identifies the nodes in the graph and connects the appropriate ports. In this example case, the resulting structural variation is an interchromosomal translocation.

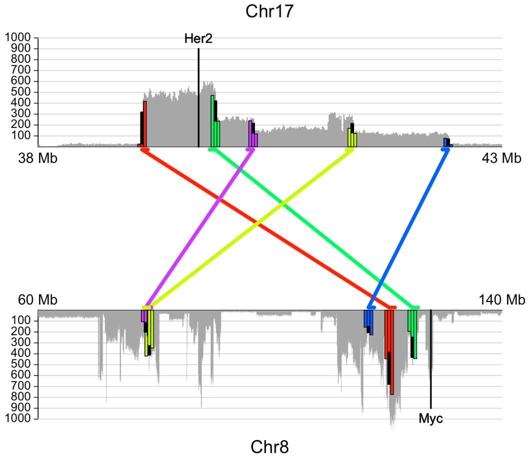
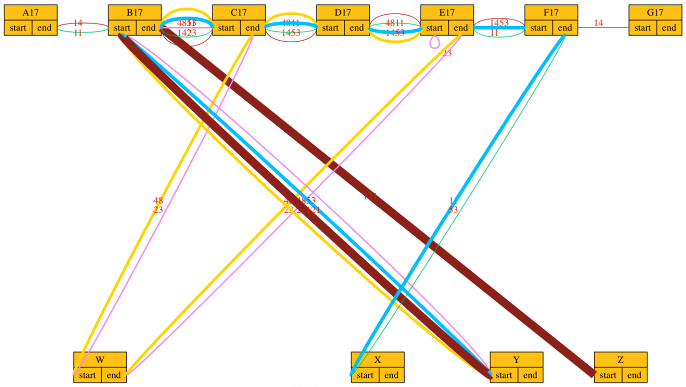
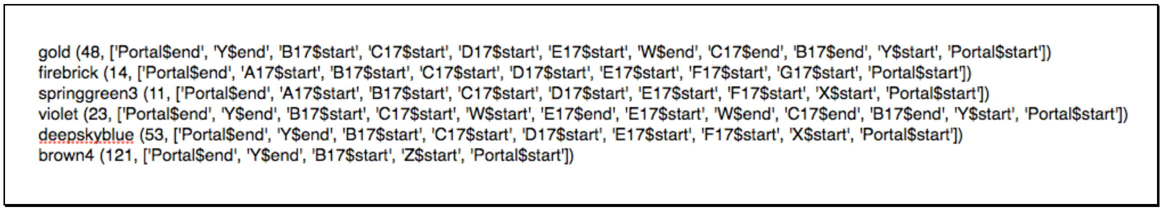


fig 5. Above: diagram of the structural variations in the HER2 region, created before my arrival. Below: A diagram of the HER2 region using my program. The top row of nodes are DNA fragments from Chr17, while fragments from Chr8 make up the bottom row. The algorithm has highlighted three maximum capacity paths. Below is a legend with the max capacities and their corresponding paths. Nodes are represented as “Node$port”.



DNA Fragments Mapping



```
gold (48, ['Portal$end', 'Y$end', 'B17$start', 'C17$start', 'D17$start', 'E17$start', 'W$end', 'C17$end', 'B17$end', 'Y$start', 'Portal$start'])
firebrick (14, ['Portal$end', 'A17$start', 'B17$start', 'C17$start', 'D17$start', 'E17$start', 'F17$start', 'G17$start', 'Portal$start'])
springgreen3 (11, ['Portal$end', 'A17$start', 'B17$start', 'C17$start', 'D17$start', 'E17$start', 'F17$start', 'X$start', 'Portal$start'])
violet (23, ['Portal$end', 'Y$end', 'B17$start', 'C17$start', 'W$start', 'E17$end', 'E17$start', 'W$end', 'C17$end', 'B17$end', 'Y$start', 'Portal$start'])
deepskyblue (53, ['Portal$end', 'Y$end', 'B17$start', 'C17$start', 'D17$start', 'E17$start', 'F17$start', 'X$start', 'Portal$start'])
brown4 (121, ['Portal$end', 'Y$end', 'B17$start', 'Z$start', 'Portal$start'])
```

By tracing through the Her2 region, the algorithm has predicted which sequences of the genome have been strung together and in which order. This has implications for tracing gene fusions, some of which we have observed to result from a series of 2 or more translocations. When translocations are close together, as we observe in chromosomes 8 of SK-BR-3 especially, it will be unclear which enhancers are close to which genes. By choosing the most likely actual paths of sequences in the genome, we can better determine the post-mutation distance between enhancers and genes, as well as gain a list of variants to see the final picture of the cancer genome after the mutations have taken place. We are now starting to apply the algorithm on a genome-wide scale in SK-BR-3. The next step will be to apply heuristics and determine the best strategy for optimization. The ability to recreate a complex graph of the HER2 region, gain information about different DNA fragments, and visualize the structural variations will become a powerful tool for retracing the process of cancer genome mutation.

Discussion

Many graph models exist already, but currently there are none that adequately depict DNA fragment paths. Our design, where nodes possess "start" and "end" ports, allows for the possibility of entering a node in two ways. If an edge connects to the end port, then the DNA fragment will be read back to front. This feature is crucial for an algorithm that detects structural variations. It will also be a useful tool for any other research involving rearranged genomes. My algorithm also stores DNA code information, and once the correct mapping has been found, it can be made to print out the full DNA sequence of that mapping. Currently the algorithm has only retraced the HER2 region, but ultimately it could be used to analyze the whole SK-BR-3 genome. A full mapping of the SK-BR-3 genome and the processes of structural variation within it will serve as an invaluable resource for breast cancer researchers, allowing them to better understand the overall transformations in cancer genomes and, more specifically, how these changes cause HER2 amplification. When the Schatz lab publishes their findings, my project will be the main cited source for this new discovery, and a footnote will reference my final report, explaining the methods and findings. Ultimately the information gained from this project will be applied to future research on breast cancer tumors to help defeat this devastating disease.

Acknowledgements

Many thanks to Maria Nattestad, my graduate student mentor, and Michael Schatz, my principal investigator, for their patience, guidance, and belief in me over these 10 weeks. I also thank the URP program directors Michael Schatz, Anne Churchland, and Kim Creteur. My research experience was funded, in part, by the National Science Foundation award DBI-1350041 to MCS.

References

Baker, Monya. "Structural Variation: The Genome's Hidden Architecture." *Nature Methods Nat Meth* 9.2 (2012): 133-37.

Chaisson, Mark J. P., et al. "Resolving the Complexity of the Human Genome Using Single-molecule Sequencing." *Nature* 517.7536 (2014): 608-11.

Navin, Nicholas, et al. "Tumor Evolution Inferred by Single-cell Sequencing." *Nature* 472 (2011): 90-94.

Roberts, Richard J., Mauricio O. Carneiro, and Michael C. Schatz. "The Advantages of SMRT Sequencing." *Genome Biol Genome Biology* 14.6 (2013): 405-09.