Modeling of Linker Stoichiometry for Optimization of DNA Nanostructure Self-Assembly

Maddie Thomas
Department of Computer Science
Simpson College
Indianola, Iowa 50125
maddie.thomas@my.simpson.edu

Abstract

The Chemistry Department at Simpson College is constructing complex nanostructures by linking together many triangle shaped DNA nanostructures. The experiments conducted by the Chemistry Department have resulted in low yields of connected triangles, which led us to believe that the ratio of linkers to triangles is important. Our program simulates the assembly process using thermodynamics of DNA hybridization. The program outputs the final component counts, the most important of which is the number of good connections. Currently we know that the ratio is best at one linker for every triangle pair. Most importantly, we have determined that the curve is steeply dependent on the ratio, meaning the cause for inconsistencies in experiments is a result of a tiny margin for error in dispensing the components for the reaction.

1 Introduction

This project was in Association with the Chemistry Department at Simpson. In the Fall Semester of 2014 we began a project for our Introduction to Algorithms Class. The project we were presented with was to create an algorithm to solve a problem related to DNA computing. DNA computing is a new area of research where DNA strands are used as the substrate for carrying out the computation. DNA molecules consist of two complimentary strands of nucleotides bound by the Watson-Crick pairing in a helix structure. Single strands of nucleotides can be created artificially, and woven in various shapes using the complimentary properties of the nucleotides. Such shapes are known as DNA origami. They can be used further as building material in nanotechnology. This motivates the research done in design and construction of DNA origami.

At Simpson, the Chemistry department has been working on DNA origami and they have are using a design that employs triangular shapes as building blocks for more complex DNA origami constructs. The following figure shows a basic visual of the Tri Origami Assembly.

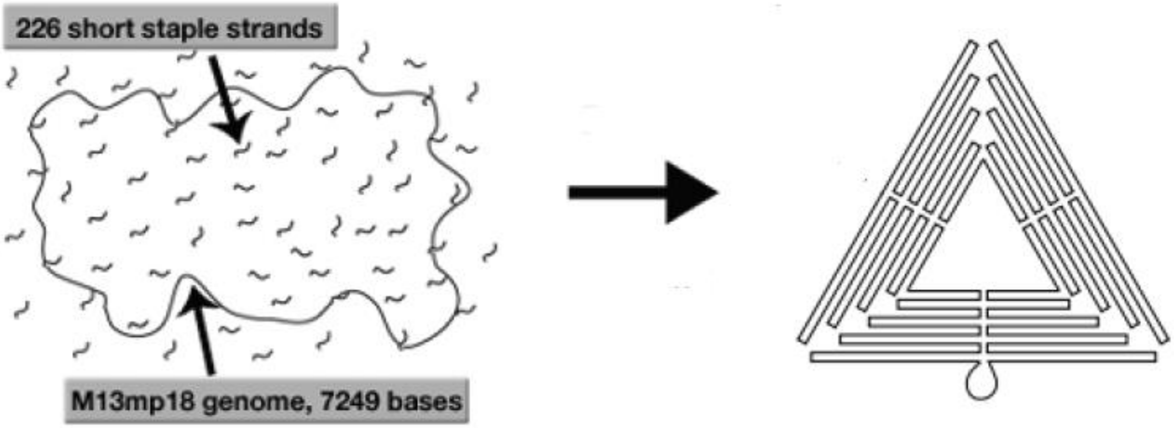

Figure 1: Basic visual of what goes into the assembly of a Tri Origami structure. DNA origami structures are created from the M13 viral genome and hundreds of short staple strands. The staple strands fold the M13 genome into the desired shape. Any shape can be created from this process; the triangle used in this study is shown.

The basic idea in this design is to attach a short single stranded linker to the sides of the triangles to bind triangles together. These short nucleotide sequences will naturally bind together with other nucleotide sequences that are complimentary in a process called annealing. Annealing is a thermal process where DNA is cooled slowly from a high temperature, allowing the correct connections to form. The overall process is called self-assembly. Self-assembly is a process by which components may form complex structures autonomously. Due to the natural properties of DNA structures, self-assembly will occur when DNA molecules are allowed to interact in favorable conditions. By linking together many triangle shaped DNA nanostructures, they are able to construct more complex nanostructures.

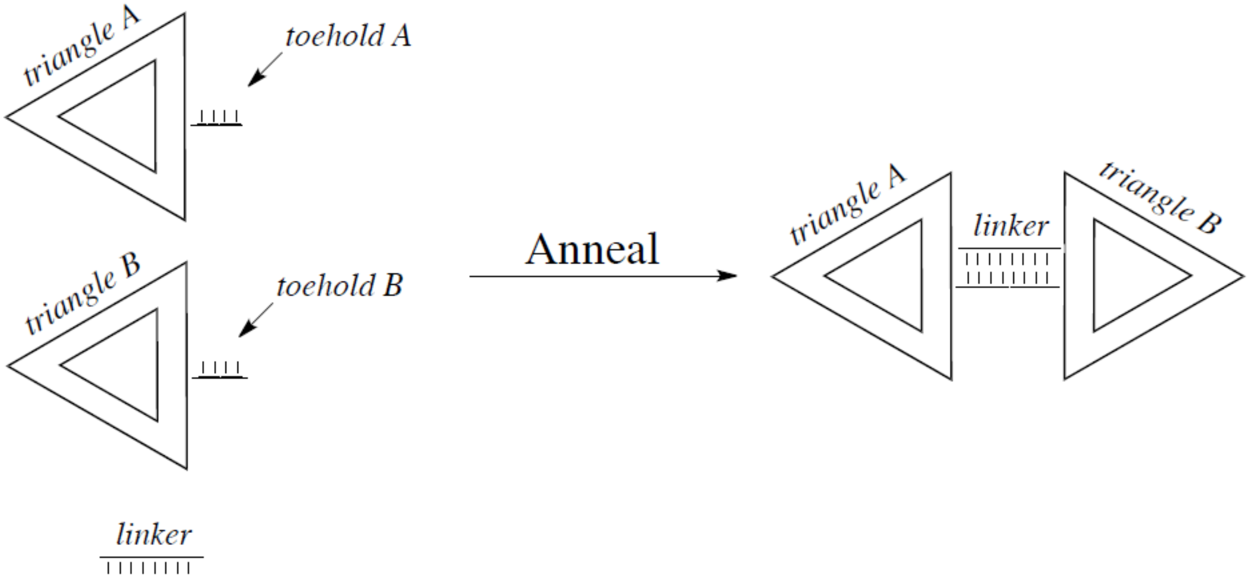

Figure 2: The linker is a unique strand that will link together triangle A and triangle B. One half of the linker is complimentary to toehold A and the other half is complimentary to toehold B.

Each DNA linker has two halves. One half of the linker is complimentary to a certain triangles' toehold and the other half is complimentary to another specific triangle. Therefore, a linker brings together two specific triangles. In this fashion, complex structures can be constructed by matching up the proper triangles. In order to provide a better chance of linkers finding their appropriate triangles, multiple identical linkers are put into the solution. Here is where a problem arises. If two triangles, that are meant to be matched together, both receive a different linker, since more than one identical linker is in the solution, then their connection is blocked and this will reduce the overall yield of connected triangles. The experiments conducted by the Chemistry department have resulted in low yields of connected triangles, which led us to believe that the ratio of linkers to triangles is important.

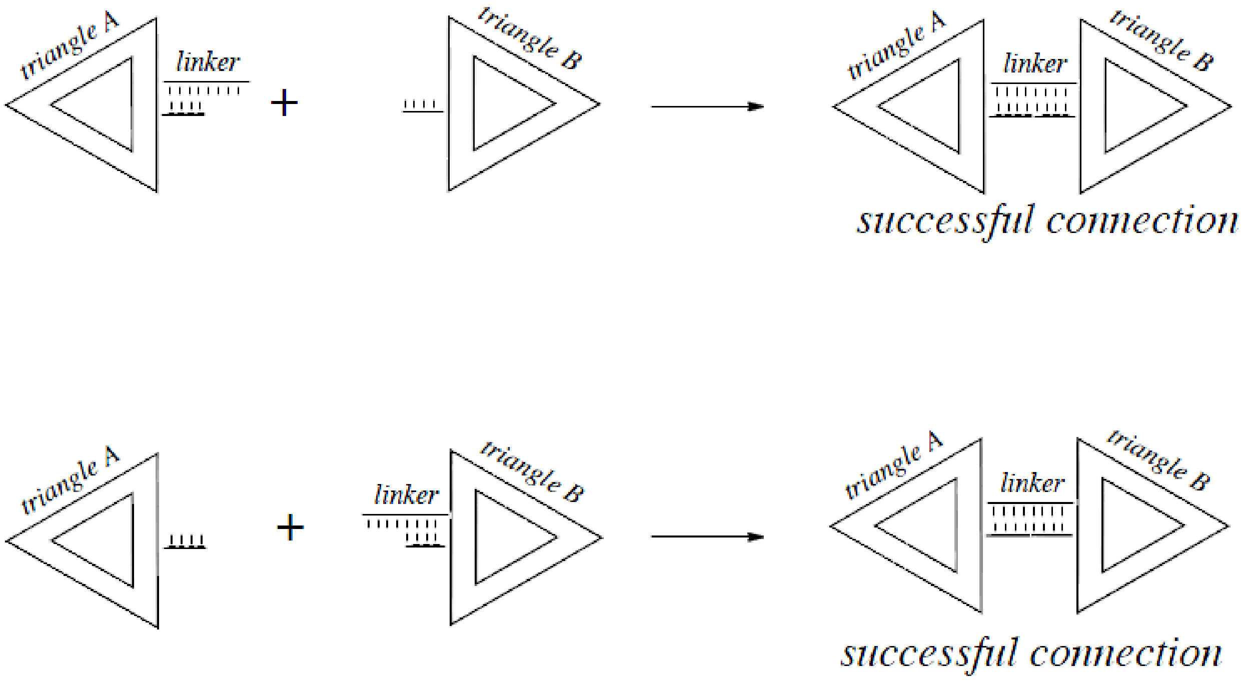

Figure 3: One half of the linker is complimentary to a certain triangles' toehold and the other half is complimentary to another specific triangle. Therefore, a linker brings together two specific triangles. These are examples of successful connections.

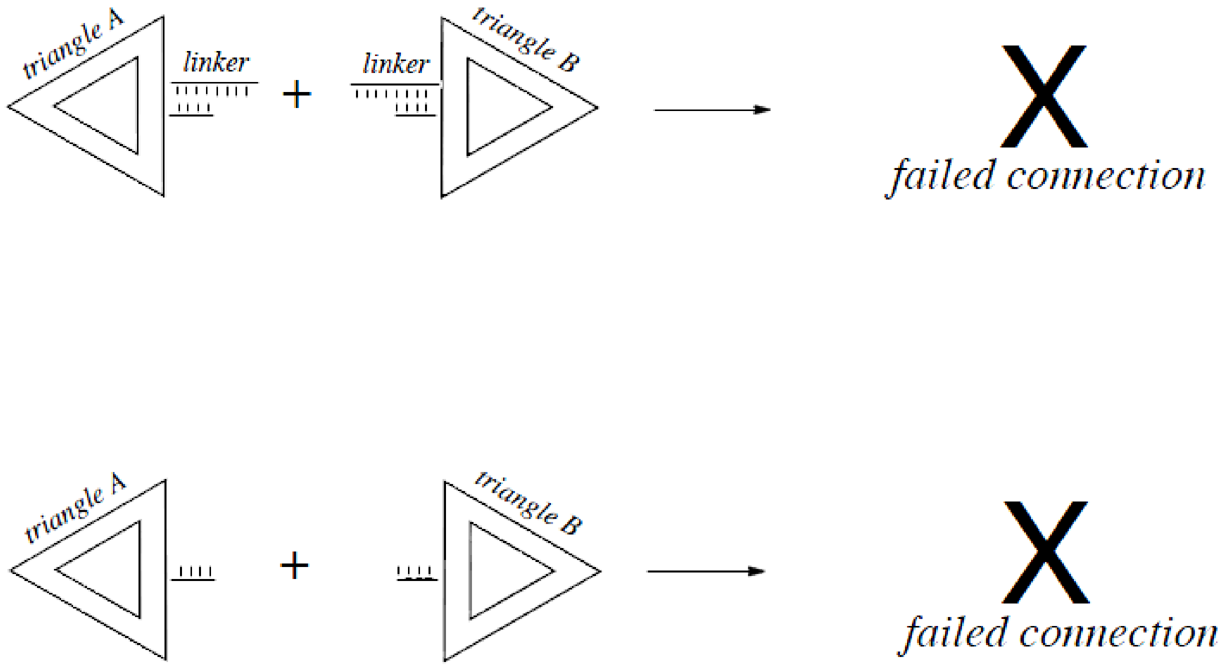

Figure 4: *(top)* The triangles both receive the linker that would connect them together, however, since they both received the linker, no connection can form. *(bottom)* Neither of the triangles intended to be linked received a linker, no connection can form.

So essentially, the question we were presented with is how to choose the ratio of linkers to triangles so that in the DNA solution the majority of the triangles bind correctly to form the desired shape. Our solution to this problem is to simulate the assembly of these nanostructures with varying ratios of linkers to triangles in order to find the optimum ratio for producing a high number of good connections.

2 Algorithm

One of the first things we needed to do was replicate the curves of the melting temperatures of the two DNA sequences that were being used. To do this, we used an equation:

$$f = \frac{1 + C_T e^{[\frac{\Delta S°}{R} - \frac{\Delta H°}{RT}]} - \sqrt{1 + 2C_T e^{[\frac{\Delta S°}{R} - \frac{\Delta H°}{RT}]}}}{C_T e^{[\frac{\Delta S°}{R} - \frac{\Delta H°}{RT}]}}$$

What this equation does is simulate the melting of DNA and will give us a fraction of double stranded DNA at some temperature, T, DNA concentration, C_T, and DNA sequence, $\Delta S°$ & $\Delta H°$. In this equation R is a gas constant. The reason this is important is because it will tell us what percentage of the triangles and linkers have been connected at any given temperature. This will allow us to calculate and check for new connections made when the temperature changes. We call these new connections events.

Since we are working with two different DNA sequences that each have a different annealing temperature, there are actually two sets of events occurring, $f1$ events, which refer to the triangle with the DNA sequence that begins to anneal at a higher temperature, and $f2$ events, which refer to the triangle with the DNA sequence that begins to anneal at a lower temperature. We can then calculate the current and previous $f1$ & $f2$ values at each temperature step then subtract to find the number of both $f1$ & $f2$ events that occurred from the previous temperature to the current temperature. This means that at each temperature iteration of the program, the equation will allow us to calculate the amount of events that occurred for both $f1$ & $f2$ events. Then we are capable of allocating each event into a classification.

Once we had the melting curves replicated within the program, we then needed to determine all the events that could occur with the different components in the solution so that we could account for them in the program. There are three original components in the solution and then three types of components that could be formed by an event. The original components are the ones introduced by the experimenter. These would be the linker, which is labeled $p0$ in the program, the triangle that anneals at a higher temperature, which is labeled $p1$ in the program, and the triangle that anneals at a lower temperature, which is labeled $p2$ in the program.

Since the linkers, $p0$, are supposed to connect to their $p1$ triangle location first, a linker connected with only a $p1$ triangle is a good connection because there is a high chance that this connection will also connect with a $p2$ triangle later in the simulation to create a final good connection. In the program this is represented by the variable Pg, for number of good connections. The value of this variable in the program should go up as the assembly begins and then down once final good connections begin to be made.

The next classification is when a linker connects with only a $p2$ triangle. This is classified as a bad connection, because it is very unlikely that this connection will turn into a final good connection by adding a $p1$ triangle. In the program this is represented by the variable Pb, for number of bad connections. The value of this variable in the program, in an optimum case, should stay as low as possible. The value could go up at the beginning and then go down slightly at the end due to the probability of a $p1$ triangle connecting to make a final good connection.

The third classification is one that has already been referenced, final good connection. When a linker connects with both a $p1$ triangle and a $p2$ triangle, this is a final good connection. In the program this is represented by the variable Pgg, for number of final good connections. The value of this variable in the program should start increasing as the temperature approaches the annealing temperature for the lower annealing triangle, and as this number goes up, the Pg number goes down. This is the variable that we are the most concerned with as it represents the overall number of complete good connections.

Once we had the events classified, we needed to figure out a way to allocate the events in such a way as to simulate what would happen in the solution. Since all of these events occur quite randomly in the solution, it is hard to say what will happen in what situation. We know that an $f1$ event can be either $A: p0 + p1 = Pg$ or $B: Pb + p1 = Pgg$ and that an $f2$ event can be either $A: p0 + p2 = Pb$ or $B: Pg + p2 = Pgg$. We decided that the best way to allocate events would be by a biased random chance. The bias is calculated by using the remaining amounts of $p0$, $p1$, $p2$, Pb, and Pg in the solution. For the $f1$ events the bias is based on the remaining amounts of $p0$ and Pb and for the $f2$ events the bias is based on the remaining amounts of $p0$ and Pg. This biasing is accurate because it is unlikely for Pb to combine with a $p1$ triangle because the annealing phase for a $p1$ triangle has most likely passed. Likewise it should be less likely for a $p2$ triangle to combine with a linker, $p0$, because the annealing phase is underway for the $p1$ triangle and there should be more Pg's available for combination.

For an overview, the simulation iterates through the temperature decrease from $100°C\ to\ 0°C$, which is actually in Kelvin in the program. Currently the program iterates by one degree but could go by any amount. The smaller the degree change, the smaller the number of events allocated at each step. For each temperature step the current and previous values of $f1$ are calculated and the difference is found. This difference is the percentage of $f1$ events that need to be allocated. We can then find the number of $f1$ events that occurred. The current and previous values of $f2$ are calculated and the difference is found. This difference is the percentage of $f2$ events that need to be allocated. We can then find the number of $f2$ events that occurred. Then, those event counts are passed into a function that updates all the P values which includes $p0$, $p1$, $p2$, Pb, Pg, and Pgg. This function is where the bias that was described earlier is applied and the events are allocated to certain classifications, updating the counts of all P values along the way. The program finishes once $0°C$ is reached.

3 Results

The program at present is very versatile, as it has not been finished out with a UI. As such we can change how many simulations we do at any time and are capable of simulating a large amount of assemblies in little time. This means we can simulate many assemblies all at once and then do further manipulations to interpret the results.

The following figure depicts the results from a large run of the simulator that was conducted. For this multiple simulation run, the constants were the starting values of $p1$ and $p2$ and the DNA sequences used. Both $p1$ and $p2$ started at 1000. The variable that we altered each time was the starting value of the linkers, $p0$. This was changed in accordance with the ratios we were testing. For example, a ratio of 10 refers to 10 $p0$ linkers for each $p1$ triangle.

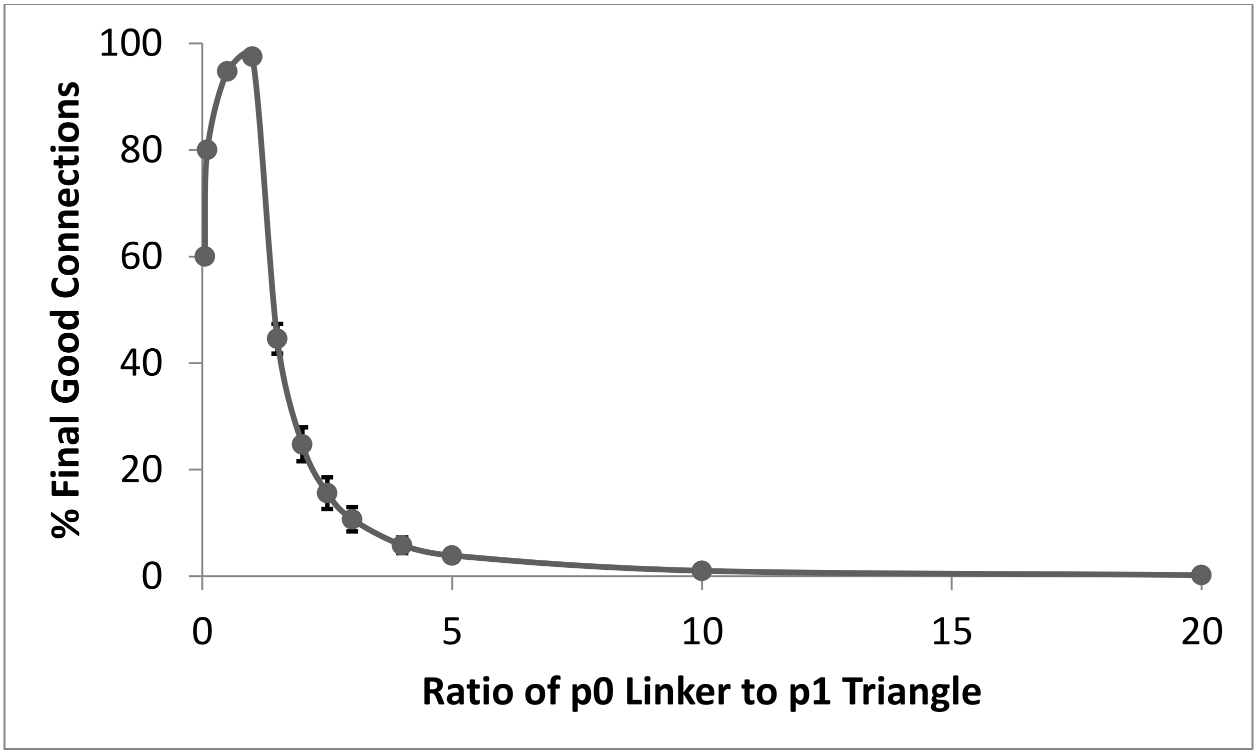

Figure 5: Simulation of thermal assembly of a DNA nanostructure from $100°C\ to\ 0°C$. The ratio of $p0$ linker to $p1$ triangle was varied. For example, a ratio of 10 refers to 10 $p0$ linkers for each $p1$ triangle. There are an equal number of $p1$ triangles and $p2$ triangles in each simulation. We plotted the percentage of good connections at the completion of the simulations, which came from the value Pgg . Each point represents the average of 100 simulations. The error bars represent one standard deviation. The value peaked at a ratio of 1.

The figure above is extremely significant. Using this data it is easy to see that the final number of good connections, Pgg, peaks at a ratio of 1. This is important because any higher of a ratio and the yield drops off steeply. What is interesting is what happens with a lower ratio. It was surprising, although consistent with previous experiments, that the yield stayed pretty high for ratios less than 1. What a ratio less than 1 means is that there are extra $p1$ and $p2$ triangles for each linker that is intended to connect them. This is an accurate projection of these ratios because this eliminates the problem of identical linkers blocking the connection of the two triangles, so although there may not one hundred percent yield because some of the linkers don't find their triangles, there are no extra linkers to block connections so when connections can occur, they do.

One of the most important findings in this research and the creation of this algorithm is the margin for error. In the lab, two seemingly identical experiments would be conducted and the results would be vastly different. By using this data, we can see that the inaccuracy of lab equipment, such as pipets, can actually cause these low yields. This is because the intended ratio may be 1, but the steep curve means that any deviation from the exact ratio can be detrimental to yields, especially deviating towards a higher ratio.

4 Future Work

The main goal for this research to continue towards is better fit to real data from the lab. More experiments will be conducted in order to compare more thoroughly the result set from the simulation. Currently the data that we have simulated is very accurate for lower ratios, but as the ratios climb higher, such as a ratio of 20, the simulation results differ from lab results. Clearly there is further research to do in this particular area of the simulation.

In the future, we would also like to implement all four toeholds into the simulation, instead of just one. This is because that is what is actually happening in the lab and would most probably help us to make our simulation even more precise in the smaller ratios. This implementation could also help to fix the inaccuracies in the higher ratios as well.