Rithvik Arun
Jacob Hreshchyshyn

MAT 243 Project 3
Sequencing

In this project our goal was to determine the entire original RNA sequence from two sets of fragments. This meant that we had to find a sequence that would satisfy both enzymes when they were applied to the RNA chain which would result in the proper fragments being created. The fragments were created by two enzymes that broke the 12-link RNA chain. The first enzyme broke the RNA chain after each G link was applied to a 12-link chain. The fragments obtained were AC, UG, and ACG. The second enzyme broke the RNA chain after a C or U link which resulted in the fragments being U, GU, and GAC. Since this is a 12 link RNA chain it means that there are 12 bases or spots (See Figure 1). For organization we will be calling the first enzyme, enzyme A, and the second enzyme, enzyme B.

Figure 1: Empty RNA Sequence

Now we will explain how the RNA chain breaks after Enzyme A and Enzyme B. Enzyme A breaks the chain after every G link. This means that everytime there is a G in the sequence there is a break. In order to see this more clearly, start from the beginning of the sequence and highlight each base with one color until you see a G (include the G). Once you have seen a G highlight the next base with a different color and keep following this pattern. Once you get to the end of the sequence you will find that all sets that you have highlighted are actually fragments of enzyme A (See Figure 2).

U G A C G U G A C G A C

Figure 2: This is an example RNA sequence that we came up with. The different colors show the the fragments obtained after applying Enzyme A. As you can see some of the colors are the same which means that there are duplicate fragments in the sequence. Also notice that all the fragments are valid with Enzyme A.

Enzyme B works similarly. Enzyme B breaks the chain after every U or C link. This means that everytime there is a U or C in the sequence there is a break. In order to see this more clearly, start from the beginning of the sequence and highlight each base with one color until you see a U or C (include the U or C). Once you have seen a U or C highlight the next base with a different color and keep following this pattern. Once you get to the end of the sequence you will find that all sets that you have highlighted are actually fragments of enzyme B (See Figure 3).

U G A C G U G A C G A C

Figure 3: We are using the same sequence as Figure 2. The different colors show the the fragments obtained after applying Enzyme B. Also notice that all the fragments are valid with Enzyme B.

Now that we know how the RNA chain breaks when Enzyme A and Enzyme B are applied, we can form our sequence. To start we decided to look at the beginning of the sequence. We noticed that the sequence must work when both Enzyme A and Enzyme B are applied. This meant that when the RNA chain was broken in the beginning, fragments should be valid to specific enzymes that were used. Due to this we decided to try each fragment from both enzymes at the beginning of the sequence. We found that the sequence had no chance of starting with a G because that would break the sequence and give us the fragment G which is not a valid fragment for enzyme A (See Figure 4). This eliminated GU and GAC.

G U _ _ _ _ _ _ _ _ _ _

G A C _ _ _ _ _ _ _ _ _

Figure 4: GU and GAC will not be valid in the beginning of the sequence because it will leave us with the fragment G (highlighted in blue) which is not a valid fragment of enzyme A.

We also found that there was no chance of the sequence starting with AC or ACG because when this was broken using enzyme B it would leave us with the fragment AC which is not a fragment for enzyme B (See Figure 5).

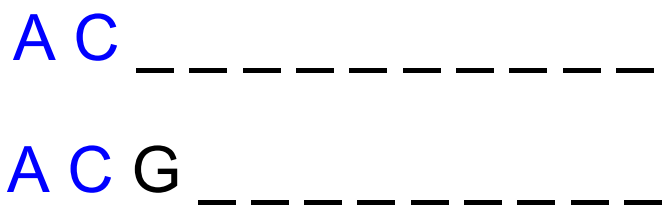

Figure 5: After applying enzyme B the link would break to give us the fragments AC which is not a valid fragment for enzyme B.

This left us with the fragments U and UG. We found that U will work with enzyme B but UG will work with both enzyme A and B (See Figure 6).

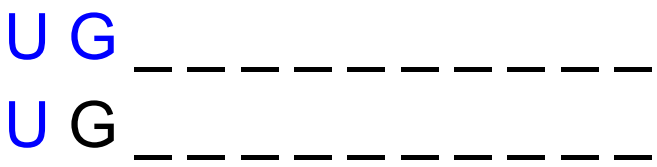

Figure 6: The first sequence shows the break after enzyme A which yields the fragment UG which is a valid fragment of enzyme A and the second sequence shows the break after enzyme B which yields the fragment U which is valid fragment of enzyme B.

Due to both enzymes breaking valid fragments we know that the beginning of the sequence is UG.

Next we can move on and try to find the end of the sequence. We tried to find the end of the sequence similar to how we found the beginning. We tried each fragment from both enzymes to see which would yield a valid break for both enzymes, but we added a G or U/C before each fragment because the fragments follow either a G link or U or C link

depending on the enzyme used. This left us with nine combinations which included GUG, GACG, and GAC for enzyme A and UU, CU, UGU, CGU, UGAC, and CGAC for the enzyme B. We found that the combinations that would work for both enzymes were UGAC and CGAC. Both would give us GAC for enzyme B which is a valid fragment for the enzyme and would give AC for enzyme A which is a valid fragment of enzyme A (See Figure 7).

U G _ _ _ _ _ _ U G A C Enzyme A
U G _ _ _ _ _ _ U G A C Enzyme B
U G _ _ _ _ _ _ C G A C Enzyme A
U G _ _ _ _ _ _ C G A C Enzyme B

Figure 7: UGAC and CGAC will give us AC for enzyme A which is a valid fragment for enzyme A and GAC for enzyme B which is a valid fragment for enzyme B.

As we tried the other combinations none of them satisfied both enzyme A and enzyme B. UGAC and CGAC are the only combinations that worked for both enzymes and therefore the ending of the sequence must be AC. Because either U or C can precede GAC, we know that the ending of the sequence must be GAC (See Figure 8).

U G _ _ _ _ _ _ _ G A C

Figure 8: The RNA sequence so far

We have our beginning and end and now we must focus on the elements in the middle. As we were coming up with the fragments that could fill up the bases in the middle we

found that there are multiple ways the fragments can be placed that will satisfy both enzymes (See Figure 9).

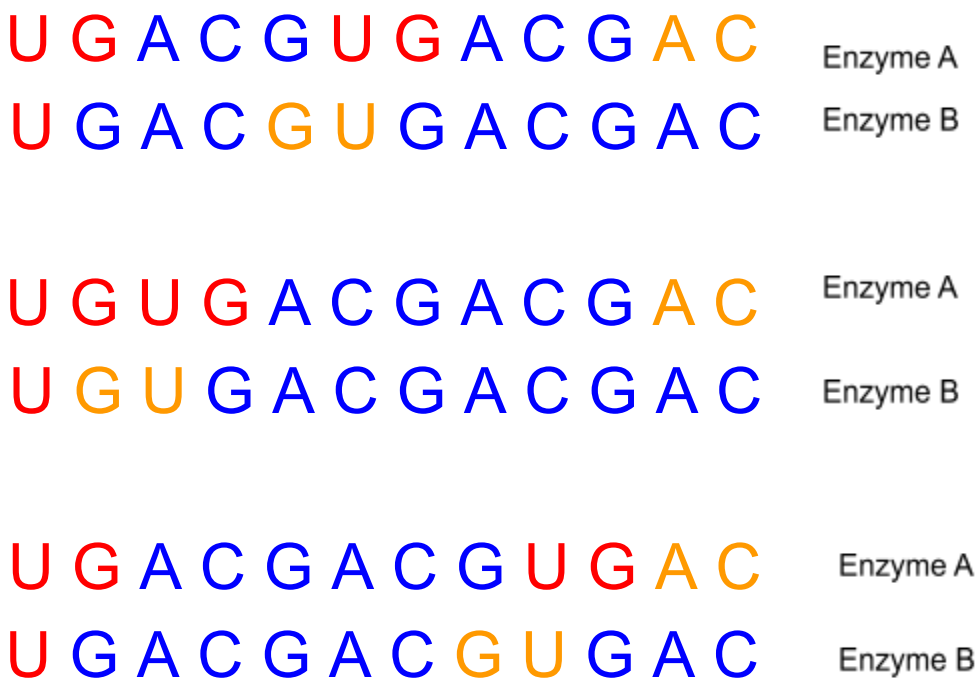

Figure 9: These are three combinations that we came up with that all satisfy both enzymes. As you can see we have added both enzymes to all the three sequences and they all yield fragments that are specific to the enzyme. The different colors represent the different fragments. Notice that there are some duplicates in the sequences which are used by the same color and how all the sequences follow with the UG_ _ _ _ _ _ _ _ _ GAC format.

Therefore we cannot determine the entire original sequence from the two sets of fragments. Multiple sequences give us multiple RNA chains which makes us unable to determine the correct sequence from the two sets of fragments. There is no way to tell which sequence is “THE” sequence. Although knowing how many of each fragments we have may help, there could still be multiple combinations that could be formed. This is because there must always be duplicates of fragments to complete the RNA chain and this allows fragment duplicates to be reordered, allowing multiple chains to be formed. As a result, it wouldn't be possible to determine which RNA sequence is the correct one. Using the information given, UG_ _ _ _ _ _ _ GAC is the most we can interpret.