Grand Valley State University
ScholarWorks@GVSU

Student Summer Scholars Undergraduate Research and Creative Practice

9-2018

The Combinatorics of the Foldings of RNA

Katrina Teunis
Grand Valley State University, teuniska@mail.gvsu.edu

Follow this and additional works at: https://scholarworks.gvsu.edu/sss

Part of the Genetics and Genomics Commons

Recommended Citation
Teunis, Katrina, "The Combinatorics of the Foldings of RNA" (2018). *Student Summer Scholars*. 195.
https://scholarworks.gvsu.edu/sss/195

This Open Access is brought to you for free and open access by the Undergraduate Research and Creative Practice at ScholarWorks@GVSU. It has been accepted for inclusion in Student Summer Scholars by an authorized administrator of ScholarWorks@GVSU. For more information, please contact scholarworks@gvsu.edu.

The Combinatorics of the Foldings of RNA

Katrina Teunis
Advisor Dr. Lauren Keough

Abstract

RNA, much like DNA, is made up of four building blocks called nucleotides, Adenine, Guanine, Cytosine, and Uracil. These nucleotides form words that like to fold in on itself and bond together, each type of nucleotide bonding with only one other type of nucleotide. Therefore, order and number of nucleotides present will determine how many times the strand of RNA can fold. Using these guidelines, we considered what happens when we have only one bonding pair. Expanding on what was proven in "k-Foldability of Words" (2017), we were able to expand on the number of ways a word can fold by adding to the list of ways any word of length $2n$ can fold. We also approached the problem from a different view by looking at how words with the same length and foldability compare to each other and defining operations between these words.

1 Introduction

Ribonucleic acid or RNA is a chain of nucleotides, much like deoxyribonucleic acid or DNA, that aid in the coding and decoding of genes. Similar to DNA, RNA has four nucleotides, specifically uracil (instead of DNA's Thymine), cytosine, adenine, and guanine. We represent each nucleotide with its first letter, U, C, A, and G respectively. Unlike DNA's typical double stranded helix formation, RNA is most often found as a single strand folded in on itself. Because the RNA nucleotides do not have a second strand to bond with, they are inherently less stable, causing it to fold and bond with itself; A bonds with U and C bonds with G. We call letters that bond with each other complementary pairs.

Consider words of letters A, U, C, and G like $AUCGAUAC$. How many ways such a word of RNA can fold is determined by the order in which the nucleotides are arranged. For example, the word $AUAU$ can fold two ways, as shown in Figure 1.

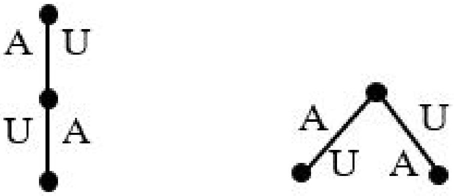

Figure 1: The two ways of folding the word AUAU.

We represent the ways a word can fold in Figure 1 by "wrapping" the word around a graph theoretic tree. Here, each nucleotide is a side of an edge so a "complementary pair" (or a pair of bounded nucleotides) sit on opposite sides of the same edge.

Another way to represent this bonding is through "non-crossing matchings". In Figure 2 are the non-crossing matchings associated to the two ways that $AUAU$ can be folded, as shown in Figure 1.

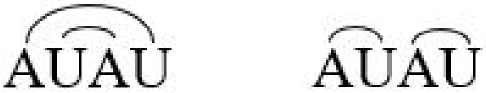

Figure 2: The non-crossing matchings of the two foldings of AUAU.

In the non-crossing matching representations, we show the bonded pairs by drawing an arc connecting the two letters that are bonded (or are on two sides of the same edge of a tree). From this representation we note that the edges cannot cross because this would not be a proper folding around a tree, and is not possible in real RNA. The non-crossing matching representation helps us see that not all words with the same number of letters fold in the same number of ways. For example, the word AAUU can only fold one way because the bonds cannot cross each other (see Figure 3).

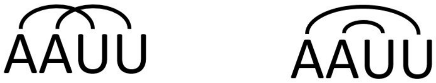

Figure 3: To the left is an illegal non-crossing matching
and on the right is a legal non-crossing matching.

While this project is inspired by the foldings of RNA we generalized this ideas to apply to a broader collection of concepts. This idea was first introduced by Black, Drellich, and Tymoczko in 2015 ([3]). They generalized to account for any length of word with any number of different complementary pairs. So, instead of four nucleotides A, U, C, G we have n complementary pairs A_1, B_1, A_2, B_2, ... A_n, B_n, where A_i and B_i are complementary pairs. In 2017 a research group, submitted [2] a related paper exploring the number of ways various words can fold. The authors proved results about the number of ways words with only one letter A and complement $\bar{A}$ can fold. Similarly, we explored the properties of words with only one letter and complement, calling them A and B.

The overall interest in this area of combinatorics has been on how many ways a word of a certain length can fold. In the past a folding has been defined as the number of different trees a word can be wrapped around, however here we will define is as follows.

Definition. A word *folds* if there is a legal non-crossing matching of all of the letters and complements.

Black, Drellich, and Tymoczko [3] established an upper bound by proving that the greatest number of ways a word of length $2n$ can fold is then nth Catalan number, C_n, and the next greatest is C_{n-1}. To this knowledge the research group added in [2] that a word of length $2n$ could also fold $0, 1, 2, ...n$ ways. However, we found that if the word is specified to have an equal number of letters and complements it cannot fold zero ways.

Definition. Define *adjacent pair* as one A and one B that are next to each other in the word.

Definition. Define *length* as $\ell = 2n$ where n is the number of As in the word.

Proposition 1.1. *If a word has length $\ell \geq 2$, is made of only one letter and its complement, and there is an equal number of each, then the word will be foldable.*

Proof. Given a word with length $\ell = 2n$ that is made of n letters, A, and n of its complement B. To prove that something is foldable we must prove that it can be grouped into pairs of one A and one B. To do this we will use strong induction.

We will begin with our base step $n = 1$. This there is only one word that has $n = 1$ and that is AB. Since this word only has one A and one B, it is already paired and we are done.

Next we will do our inductive step. We will assume as our inductive hypothesis that cases $n = 1, 2, 3...k$ can be paired and we will show that case $n = k + 1$ can be paired. Let W be a word with $n = k + 1$. Because there are both As and Bs in this word, there has to be at least one adjacent pair. Pair these together and remove them from the word, we can do this because, since there are next to each other, removing them won't effect the foldability of the word. After this adjacent pair is removed the length of this word is now $\ell = 2(k + 1) - 2$ or $l = 2k$, this means that $n = k$ and we know from our inductive hypothesis that any word with $n = k$ can be paired.

Therefore, by the principle of mathematical induction, any word with length $\ell \geq 2$ with equal As and Bs can be paired into groups of one A and one B, and therefore can be folded. □

Along the lines of our goal to find all the possible foldings, examining the structures of general cases will reveal possible foldings. It turns out that the word $(A)^n(B)^n$, which consists of n As and n Bs, folds only one way.

Proposition 1.2. *For all integers $n > 0$, $k((A)^n(B)^n) = 1$*

Proof. We know that in order for two letters to bond their bond must reach over exactly the same number of As and Bs, otherwise the rest of the word cannot fold. So given the word $(A)^n(B)^n$, because of the way all of the As are adjacent and all of the Bs are adjacent, each A can only bond with the B that has the same number of Bs before it as the A has after it. Since there is only one possible B fitting the criteria for every A, $(A)^n(B)^n$ can only fold once. □

2 Combining Words

Continuing in the search for ways a word of length $2n$ can fold, we specifically focused on proving (or disproving) the existence of words of length $2n$ that fold $n+1$ times. To do that we started by showing that certain words with known foldability can be combined to from bigger words with a predictable foldability.

Definition. Define *buffer letters* as the As and Bs added to the beginning and end of a word meant to increase the length without increasing the foldability.

Proposition 2.1. *Any word of the following forms, with $\ell > j$, $s > t$, and $z \geq 0$, will be $hy-foldable$ where h is the foldability of the first subword in brackets and y is the foldability of the second subword in brackets.*

- $A^z[ABA^\ell B^j A^j B^\ell]B^z[B^s A^t B^t A^s BA]$
- $A^z[A^j B^\ell ABA^\ell B^j]B^z[B^s A^t B^t A^s BA]$
- $A^z[A^j B^\ell ABA^\ell B^j]B^z[B^s A^t BAB^t A^s]$

Proof. Let k be the number of times the word $W = A^z Q B^z R$ folds, letting Q be a word with length x and is $y-foldable$ and R be a word with length s and is $h-foldable$. Because we are not altering the order of these smaller words by concatenating them, we know that W will at least have the foldability of the product of the smaller words, or $k \geq yh$.

Now let Q and R be of the form $ABA^\ell B^\ell A^j B^\ell$ with $\ell > j$ or $A^j B^\ell ABA^\ell B^j$ with $j > \ell$, because no new adjacent pairs will be formed by the joining of these subwords, the only way to get more than yt foldings is if a letter in one word and a complement in the other have an equal number of As and Bs between them. We will now show that this is not the case for any of these words. To do this we will look specifically at if the Bs in the first word can bond with any As in the second word or in the buffer letters. If this cannot happen, then neither word can bond with the other. This is because if the Bs in the first word cannot bond but the As can, then there will be Bs left unpaired, which is not possible for a valid folding.

Let ℓ, j, t, s, z be integers where $z \geq 0$, $\ell > j$ and $s > t$, also let b be the number of Bs and a be the number of As. We will look at each of the cases listed in the proposition. We will distinguish each group of As and Bs by including their exponent.

- Case 1: $A^z[ABA^\ell B^j A^j B^\ell]B^z[B^s A^t B^t A^s BA]$.

So, B in the first word has between it and A^t in the second word, $a \leq \ell + j + t - 1$ and $b \geq j + \ell + s$, remember that any buffer letters added will only increase b. So, $b \geq j + \ell + s > \ell + j + t - 1 \geq a$ meaning $b > a$.

Between B and A^s, $a \leq \ell + j + t + s - 1$ and $b \geq j + \ell + t + s$, and since $b \geq j + \ell + t + s > \ell + j + t + s - 1 \geq a$, $b > a$.

Between B in the first word and A in the second we have $a \leq \ell + j + t + s$ and $b \geq j + \ell + t + s + 1$, and since $b \geq j + \ell + t + s + 1 > \ell + j + t + s \geq a$, $b > a$.

Next, we will look B^j. So between B^j and A^t, $a \leq j + t - 1$ and $b \geq \ell + s$, this gives us $b \geq \ell + s > j + t - 1 \geq a$ or $b > a$.

Between B^j and A^s, $a \leq j+t+s-1$ and $b \geq \ell+t+s$, and since $b \geq \ell+t+s > j+t+s-1 \geq a$, $b > a$.

Between B^j and A in the last word, $a \leq j + t + s$ and $b \geq \ell + t + s + 1$, and since $b \geq \ell + t + s + 1 > j + t + s \geq a$, $b > a$.

Finally between B^ℓ and A^t, $a \leq t - 1$ and $b \geq s$, and since $b \geq s > t - 1 \geq a$, $b > a$.

Between B^ℓ and A^s, $a \leq t + s - 1$ and $b \geq s + t$, and since $b \geq s + t > t + s - 1 \geq a$, $b > a$.

Last of all between B^ℓ and A, $a \leq t+s$ and $b \geq t+s+1$, and since $b \geq t+s+1 > t+s \geq a$, $b > a$.

Therefore, because every interval between the two words had more Bs than As, no B in the first word can bond with an A in the second word. The last thing we need to check is can any of the Bs in $ABA^\ell B^j A^j B^\ell$ bond with the buffer As than could come before it, or can any of the As in $B^s A^t B^t A^s BA$ bond with the buffer Bs that come before it. In a similar fashion as above, we can see that at no point in the word $ABA^\ell B^j A^j B^\ell$ does a B have more Bs before it than As, so adding more As will not give us an equal number of each. Also there is no point in the word $B^s A^t B^t A^s BA$ than an A has more As before it than Bs, so adding more Bs will not give us an equal number of each. Therefore, there is no way to gain a new folding by concatenating these two words together.

- Case 2: $A^z[A^j B^\ell ABA^\ell B^j]B^z[B^s A^t B^t A^s BA]$.

Beginning with between B^ℓ and A^t, we have $a \leq \ell + t$ and $b \geq 1 + j + s$, and since $b \geq 1 + j + s > \ell + t \geq a$, $b > a$.

Between B^ℓ and A^s we have $a \leq \ell+t+s$ and $b \geq 1+j+s+t$, and since $b \geq 1+j+s+t > \ell + t + s \geq a$, $b > a$.

Between B^ℓ and A we have $a \leq \ell + t + s + 1$ and $b \geq 2 + j + s + t$, and since $b \geq 2 + j + s + t > \ell + t + s + 1 \geq a$, $b > a$.

Next between B and A^t we have $a \leq \ell + t - 1$ and $b \geq j + s$, and since $b \geq j + s > \ell + t - 1 \geq a$, $b > a$.

Between B and A^s we have $a \leq \ell + t + s - 1$ and $b \geq j + s + t$, and since $b \geq j + s + t > \ell + t + s - 1 \geq a$, $b > a$.

Between B and A we have $a \leq \ell+t+s$ and $b \geq j+s+t+1$, and since $b \geq j+s+t+1 > \ell + t + s \geq a$, $b > a$.

Finally between B^j and A^t, $a \leq t - 1$ and $b \geq s$, and since $b \geq s > t - 1 \geq a$, $b > a$.

Between B^j and A^s, $a \leq t + s - 1$ and $b \geq s + t$, and since $b \geq s + t > t + s - 1 \geq a$, $b > a$.

Between B^j and A, $a \leq t+s$ and $b \geq s+t+1$, and since $b \geq s+t+1 > t+s \geq a$, $b > a$.

Therefore, because every interval between the two words had more Bs than As , no B in the first word can bond with an A in the second word. The last thing we need to check is if any of the Bs in $A^j B^\ell ABA^\ell B^j$ bond with the buffer As than could come before it, or can any of the As in $B^s A^t B^t A^s BA$ bond with the buffer Bs that come before it. In a similar fashion as above, we can see that at no point in the word $A^j B^\ell ABA^\ell B^j$ does a B have more Bs before it than As, so adding more As will not give us an equal number of each. Also there is no point in the word $B^s A^t B^t A^s BA$ than an A has more As before it than Bs, so adding more Bs will not give us an equal number of each. Therefore, there is no way to gain a new folding by concatenating these two words together.

- Case 3: $A^z[A^j B^\ell ABA^\ell B^j]B^z[B^s A^t BAB^t A^s]$.

Between B^ℓ and A^t, $a \leq \ell + t$ and $b \geq j + s + 1$, and since $b \geq j + s + 1 > \ell + t \geq a$, $b > a$.

Between B^ℓ and A, $a \leq \ell+t+1$ and $b \geq j+s+2$, and since $b \geq j+s+2 > \ell+t+1 \geq a$, $b > a$.

Between B^ℓ and A^s, $a \leq \ell + t + s + 1$ and $b \geq j + s + t + 2$, and since $b \geq j + s + t + 2 > \ell + t + s + 1 \geq a$, $b > a$.

Between B and A^t, $a \leq \ell+t-1$ and $b \geq j+s$, and since $b \geq j+s > \ell+t-1 \geq a$, $b > a$.

Between B and A, $a \leq \ell+t$ and $b \geq j+s+1$, and since $b \geq j+s+1 > \ell+t \geq a$, $b > a$.

Between B and A^s, $a \leq \ell+t+s$ and $b \geq j+t+s+1$, and since $b \geq j+t+s+1 > \ell+t+s \geq a$, $b > a$.

Finally between B^j and A^t, $a \leq t-1$ and $b \geq s$, and since $b \geq s > t-1 \geq a$, $b > a$.

Between B^j and A, $a \leq t$ and $b \geq s+1$, and since $b \geq s+1 > t \geq a$, $b > a$.

Between B^j and A^s, $a \leq t+s-1$ and $b \geq s+t+1$, and since $b \geq s+t+1 > t+s-1 \geq a$, $b > a$.

Therefore, because every interval between the two words had more Bs than As, no B in the first word can bond with an A in the second word. The last thing we need to check is can any of the Bs in $A^jB^\ell ABA^\ell B^j$ bond with the buffer As than could come before it, or can any of the As in $B^sA^tBAB^tA^s$ bond with the buffer Bs that come before it. In a similar fashion as above, we can see that at no point in the word $A^jB^\ell ABA^\ell B^j$ does a B have more Bs before it than As before it, so adding more As will not give us an equal number of each. Also there is no point in the word $B^sA^tB^tA^sBA$ than an A has more As before it than Bs, so adding more Bs will not give us an equal number of each. Therefore, there is no way to gain a new folding by concatenating these two words together.

Therefore, the words $A^z[ABA^\ell B^jA^jB^\ell]B^z[B^sA^tB^tA^sBA]$, $A^z[A^jB^\ell ABA^\ell B^j]B^z[B^sA^tB^tA^sBA]$, and $A^z[A^jB^\ell ABA^\ell B^j]B^z[B^sA^tBAB^tA^s]$ can be concatenated without gaining any new foldings, meaning $k = hy$, where h is the foldability if the first bracketed subword and y is the foldability of the second bracketed subword.

□

Proposition 2.2. *Any word of the following forms, with $i > 0$, $j > \ell$, and $s > t$, have foldability xy where x is the foldability of the first bracketed word and y is the foldability of the second bracketed word.*

- $A^z[(AB)^i]B^z[(BA)^j]$
- $A^z[ABA^\ell B^jA^jB^\ell]B^z[(BA)^i]$
- $A^z[A^jB^\ell ABA^\ell B^j]B^z[(BA)^j]$

Proof. Use the same ℓ, j, a, b as in the proof of Proposition 2.1 and having $i > 0$. Since there will be no new adjacent pairs created by concatenating these smaller words, the only way for W to have a larger foldability than xy, is if there is an equal number of As and Bs between any As in Q and Bs not in Q. Now, because the word $(BA)^i$ has alternating As and Bs, we know there will not be a point where a letter can bond with any buffer letters before or after it. Also because of its alternating pattern, we only need to check if there is a single A in either $A^jB^\ell ABA^\ell B^j$ or $ABA^\ell B^jA^jB^\ell$ that has $a = b+1$ in the letters after it or $b = a+1$ in the letters before it, with or without buffer letters. We will look at each word individually.

- Case 1: $A^jB^\ell ABA^\ell B^j$

So, taking the word $A^jB^\ell ABA^\ell B^j$, distinguishing the As by their powers, and noting that all buffer letters added are Bs, we can see that A^j has $a \leq \ell+j$ $b \geq \ell+1+j$ after it, meaning $b \geq \ell+1+j > \ell+j \geq a$, or $b > a$ and that $a \neq b+1$.

Next A has $a \leq \ell$ $b \geq 1+j$ after it, meaning $b \geq 1+j > \ell \geq a$, or $b > a$ and $a \neq b+1$.

Finally, A^ℓ has $a \leq \ell - 1$ and $b \geq j$, meaning $n \geq j > \ell - 1 \geq a$, or $b > a$ and $a \neq b+1$.

Now, taking the same word we will look at the number of Bs before an A, noting that all the buffer letters added here are As, we can see that A^j has $a \geq 0$ and $b = 0$ before it meaning $b \neq a+1$.

Next A has $a \geq j$ and $b = \ell$ before it, meaning $a \geq j > \ell = b$, or $a > b$ and $b \neq a+1$.

Finally A^ℓ has $a \geq j+1$ and $b = \ell + 1$, meaning $a \geq j+1 > \ell+1 = b$, or $a > b$ and $b \neq a+1$.

Therefore words of the form $A^z[A^jB^\ell ABA^\ell B^j]B^z[(BA)^j]$ will have foldability xy where x is the foldability of $A^jB^\ell ABA^\ell B^j$ and y is the foldability of $(BA)^j$.

- Case 2: $ABA^\ell B^jA^jB^\ell$

So taking the word $ABA^\ell B^jA^jB^\ell$, noting that the only buffer letters are Bs, we can see that A has $a \leq \ell + j$ and $b \geq \ell + j + 1$ after it, and since $b \geq \ell + j + 1 > \ell + j \geq a$, $b > a$ and $a \neq b+1$.

Next A^ℓ has $a \leq \ell + j - 1$ and $b \geq j + \ell$ after it, and since $b \geq j + \ell > \ell + j - 1 \geq a$, $b > a$ and $a \neq b+1$.

Finally A^j has $a \leq j-1$ and $b \geq \ell$ after it, and since $b \geq \ell > j - 1 \geq a$, $b > a$ and $a \neq b+1$.

Now we need to look at the number of Bs before any of the As, noting that the buffer letters added here are As. We can see that A has $a \geq 0$ and $B = 0$ before it, meaning $a \geq b$ and $b \neq a+1$.

Next A^ℓ has $a \geq \ell$ and $b = 1$ before it, and since $ell > 0$, $b \neq a+1$

Finally, A^j has $a \geq \ell + 1$ and $b = j+1$ before it, and since $a \geq \ell + 1 > j + 1 = b$, $b \neq a = 1$.

Therefore words of the form $A^z[ABA^\ell B^jA^jB^\ell]B^z[(BA)^z]$ will have foldability xy where x is the foldability of $ABA^\ell B^jA^jB^\ell$ and y is the foldability of $(BA)^z$.

- Case 3: $(AB)^j$

The words we need to look at are words of the form $(AB)^j$. Because this also has the alternating pattern, it will always have $a = b$ before it and $b > a$ after it. Neither of these situations fit our conditions described above.

Therefore words of the form $A^z(AB)^jB^z(BA)^i$ will have foldability xy where x is the foldability of $(AB)^j$ and y is the foldability of $(BA)^i$.

Therefore the results in Proposition 2.1 can be extended to include Q and R of the form $(AB)^i$. □

Proposition 2.3. *In $W = A^zQB^zR$, either one or both of Q and R can be of the form $AABBAABB$ maintaining $k(W) = hy$ where h is the foldability of Q word and y is the foldability of R.*

Proof. We will be using the same a, b, j, ℓ, h, y as before. Because the word $AABBAABB$ is very similar to words of the form $(AB)^i$, we can use the results from Proposition 2.2 to show that the first A in each group cannot bond with $A^jB^\ell ABA^\ell B^j$, $ABA^\ell B^jA^jB^\ell$, or $(AB)^j$

for the same reason they cannot bond with $(AB)^i$. Using similar reasoning we can see that the second A in each group also cannot bond with $A^jB^\ell ABA^\ell B^j$, $ABA^\ell B^jA^jB^\ell$, or $(AB)^j$ because it needs $a = b+2$ any A in these words needs after it $a = b+2$. However, as we have shown in Proposition 2.2 in every case $b > a$, meaning $a \neq b+2$. Therefore $AABBAABB$ cannot bond with $A^jB^\ell ABA^\ell B^j$, $ABA^\ell B^jA^jB^\ell$, or $(AB)^j$.

We can also see that $AABBAABB$ cannot bond with itself for the same reason $(AB)^i$ cannot bond with itself. The alternating pattern causes there to always be more Bs between any A in the first word and B in the second.

Therefore, because concatenating $AABBAABB$ with $A^jB^\ell ABA^\ell B^j$, $ABA^\ell B^jA^jB^\ell$, or $(AB)^j$ does not form any new bondings Proposition 2.1 can be expanded to allow either one or both of Q and R to be $AABBAABB$ and $k(W) = hy$. □

Example. (Proposition 2.1) The word $ABAAABBAABBB$ has $n = 6$ and $k = 6$ and the word $ABAABABB$ has $n = 4$ and $k = 4$. So if $z = 13$, then the word

$$AAAAAAAAAAAAA[ABAAABBAABBB]BBBBBBBBBBBBB[BBABAABA]$$

has $n = 23$ and $k = 24$. This fits because $6 + 4 + 13 = 23$ and $4 \cdot 6 = 24$.

Example. (Proposition 2.2) The word $ABAABABB$ has $n = 4$ and $k = 4$, and the word $ABAB$ has $n = 2$ and $k = 2$. So if $z = 1$, then the word $A[ABAABABB]B[BABA]$ has $n = 7$ and $k = 8$. This fits because $4 + 2 + 1 = 7$ and $4 \cdot 2 = 8$

Example. (Proposition 2.2) The word $ABAAABBAABBB$ has $n = 6$ and $k = 6$, and the word $ABAB$ $n = 2$ and $k = 2$. So if $z = 3$ then the word $AAA[ABAAABBAABBB]BBB[BABA]$ has $n = 11$ and $k = 12$. This fits because $6 + 2 + 3 = 11$ and $6 \cdot 2 = 12$.

Example. (Proposition 2.3) The word $AABBAABB$ has $n = 4$ and $k = 3$. So, if $z = 0$ the word $[AABBAABB][BBAABBAA]$ has $n = 8$ and has $k = 9$. This fits because $4 + 4 = 8$ and $3 \cdot 3 = 9$

Example. (Proposition 2.3) The word $ABAABABB$ has $n = 4$ and $k = 4$, and the word $AABBAABB$ has $n = 4$ and $k = 3$. So if $z = 3$, then the word $AAA[ABAABABB]BBB[BBAABBAA]$ has $n = 11$ and $k = 12$. This fits because $4 + 4 + 3 = 11$ and $4 \cdot 3 = 12$

Now that we have shown it is possible to combine smaller words with known foldability to get bigger words with predictable foldability, we are ready to use that to find a word of length $2n$ with foldability $n + 1$.

Definition. Define $S_k(n, 1)$ to be all the words of length $2n$ that fold k times.

Theorem 2.4. *For all integers $n > 3$ where $n \neq p - 1$ for any prime number p. The set of words $S_{n+1}(n, 1)$ is nonempty.*

Proof. Let $n > 3$ and $z > 0$ be integers. Now take the word $w = A^zQB^R$ where $n(W)$ is one half the length of W and $k(W)$ is the foldability of W. Now let Q with $n(Q) = x$ and $k(Q) = s$ and R with $n(R) = y$ and $k(R) = t$. Both being of the form $ABA^lB^jA^jB^l$ with $l > j$ or $A^jB^lABA^lB^j$ with $j > l$. By Proposition 2.1 we know we can concatenate Q and R in W and get $n(W) = x + y + z$ and $k(W) = st$. By Proposition 4.10 in [3] we know that

Q and R have values for l and j such that $n(Q) = k(Q)$ and $n(R) = k(R)$ for all $n > 3$. So, let $n(Q) = k(Q) = x$ and $n(R) = k(R) = y$. This gives us $n(W) = x + y + z$ and $k(W) = xy$ for $x > 3$ and $y > 3$. Now let $z = x(y - 1) - y - 1$, making

$$\begin{aligned} n(W) &= x + y + z \\ &= x + y + [x(y - 1) - y - 1] \\ &= x + y + xy - x - y - 1 \\ &= xy - 1. \end{aligned}$$

Therefore, $k(W) = xy$ and $n(W) = xy - 1$, or for all $n = xy - 1$ with $x > 3$ and $y > 3$ the set words $S_{n+1}(n, 1)$ is nonempty.

However, this only works for $x, y > 3$. We can expand this idea by remembering that the word $ABAB$ has $n = 2$ and $k = 2$ and even though there is no word with $n = k = 3$ we do have $AABBAABB$ which has $n = 4$ and $k = 3$. So, we will look at the cases where $n(Q) > 3$ $k(Q) > 3$ and $n(R) = k(R) = 2$, $n(Q) = 4$ $k(Q) = 3$ and $n(R) = 4$ $k(R) = 3$, and $x > 3$ and $n(R) = 4$ $k(R) = 3$. Note: we will not be looking at $x = 2$ and $y = 2$ because $2 \cdot 2 = 4$ and $2 + 2 = 4$. We will also not be looking at $n(Q) = 4$ $k(Q) = 3$ and $n(R) = k(R) = 2$ because $2 \cdot 3 = 6$ and $2 + 4 = 6$. Neither of these results are in the set $S_{n+1}(n, 1)$.

We will first look at the case where Q is the same as defined above and $R = ABAB$, meaning $n(Q) > 3$ $k(Q) > 3$ and $n(R) = k(R) = 2$. Because $ABAB$ is of the form $(AB)^i$ we know from Proposition 2.2 that we can use the equations $k(W) = xy$ and $n(W) = xy - 1$ letting $y = 2$. This gives us $k(W) = 2x$ and $n(W) = 2x - 1$ with $z = x - 3$. So, for all $n(W) = 2x - 1$ with $x > 3$, the set words $S_{n+1}(n, 1)$ is nonempty.

Next we will look at the case where $Q = R = AABBAABB$. This means $n(Q) = 4$ $k(Q) = 3$ and $n(R) = 4$ $k(R) = 3$. We know from Proposition 2.3 that we can concatenate $AABBAABB$ with itself giving us $k(W) = 9$. Now let $z = 0$ this gives us $n(W) = 8$ which is in the set $S_{n+1}(n, 1)$.

Finally, we will look at the case where Q is the same as defined above and $R = AABBAABB$. By Proposition 2.3 know $k(W) = 3x$ and $n(W) = x + z + 4$. Now let $z = 2x - 5$, making $n(W) = 3x - 1$ which is in our set $S_{n+1}(n, 1)$.

Therefore the set of words $S_{n+1}(n, 1)$ is nonempty for all $n = xy - 1$ with

- $y > 3$ and $x > 3$
- $y = 2$ and $x > 3$
- $y = 3$ and $x > 2$

So, we can now show that there exists an $n(W)$ that is one less than $k(W)$ for $k(W) \neq p$ for all prime numbers p and $k(W) \geq 6$. This result comes from the fact that $k(W) = xy$ for the values of x and y above. By the Fundamental Theorem of Arithmetic, we know that every integer is either prime or a product of primes. So, since neither x nor y can be 1, $k(W)$ cannot be prime. However, we have shown that y can be any value greater than 1, meaning $k(W)$ can be any of the non-prime numbers if $x > 1$. So, since x has some restriction we can see that $k(W)$ cannot be $4, 5, 6$ or 7, meaning that $k(W)$ exists for all $k(W) > 7$.

Therefore, since $n(W) = k(W) - 1$, we know that the set of words $S_{n+1}(n, 1)$ is nonempty for all $n \geq 7$ as long as $n \neq p - 1$ for all prime numbers p. □

In may seem odd that n cannot be one less than a prime number, or that the words cannot fold a prime number of ways. It is not impossible for a word to fold a prime number of ways; [2] showed that there are words that fold a number of ways that is found by adding consecutive Catalan numbers, for example $C_2 = 2$, $C_3 = 5$, and $2+5 = 7$ which is prime. The method, however, which we used to build our words that folded $n+1$ ways used multiplication and therefore cut out any prime possibilities. Other than the addition of consecutive Catalan numbers, how else could a string have a prime foldability? So far we have seen that a large part of the foldability of a word is based on the Catalan numbers, either the number itself or by multiplying and adding them. If we are multiplying, then we automatically cannot produce primes by their very definition. Addition we have only seen work with consecutive Catalan numbers, so far, giving us very few primes. Finally there are the Catalan numbers themselves. However, L. T. Peabody in answering the question "Can it be shown that there are finitely many (or infinitely many) such Catalan numbers?" [1] showed that the largest prime Catalan number is $C_3 = 5$. His proof for this has been revised and simplified.

Theorem 2.5. *If $n > 3$, then C_n is not prime.*

Proof. We know that $C_n = \frac{(2n)!}{n!(n+1)!}$, or $n!(n+1)!C_n = (2n)!$. We also know that either $C_n > 2n$, $C_n = 2n$, or $C_n < 2n$. If $C_n > 2n$ then it must be some multiplicative combination of integers less than $2n$ and by definition is not prime. Similarly if $C_n = 2n$ then C_n is not prime. So, if $C_n > 2n-1$ then C_n is prime. However, since $C_n = \frac{(2n)!}{n!(n+1)!}$, C_n is prime if $\frac{(2n)!}{n!(n+1)!} > 2n-1$ or $2(2n-2)! > (n-1)!(n+1)!$. This means $2(2n-2)! - (n+1)!(n-1)! > 0$. Now, as long as $(2n-2) > (n+1)$ or $(n > 3)$, we can factor out $2(n+1)!$ to get

$$2(n+1)![(2n-2)(2n-3)...(n+2) - (n-1)(n-2)...3 \cdot 1].$$

So, if $(2n-2) > (n-1)$ and $(n+2) > 3$, $(2n-2)(2n-3)...(n+2) - (n-1)(n-2)...3 \cdot 1$ will be positive. Both of these are the case when $n > 3$. So, when $n > 3$, $2(2n-2)! > (n+1)!(n-1)!$ and C_n is not prime. Therefore $C_3 = 5$ is the largest prime Catalan number. □

3 Transforming Words

In this section we will look at operations that relate words to each other, specifically words with the same foldability. The operations we found are as follows.

Definition. A *cyclic shift* on a word W with length $2n$ is a word W' such that, for all i, $W'[i] = W[(i+j) \mod 2n]$ for any integer j.

Example. A cyclic shift of the word $ABAABB$ with $j = 1$ is $BAABBA$.

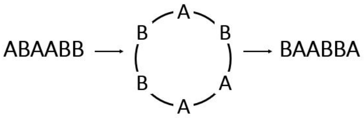

Figure 4: The cyclic shift of the word $ABAABB$ with $j = 1$

Proposition 3.1. *Given a word W, any cyclic shift of W have the same foldability as W.*

Proof. Given a word of length $2n$, by definition of their position, the first letter has no letters before it and the last letter has no letter after it. Therefore, any connection between them crosses over all of the other letters and therefore does not affect their foldability. This is the same effect as if the connection crossed over none of the other letters, which happens only when the letters are right next to each other. Therefore the first and last letters of a word can be thought of as adjacent and moved together. This will cause the word to form a circle that can be split and reshaped into a line between any two letters as shown in figure 4.

Therefore, the letter in the ith position can be moved to the $(i + j) \mod 2n^{th}$ position, or a cyclic shift, does not change the foldability of a word. □

Definition. A *reversal* is an operation on a word W of length $2n$ such that $W[i] = W[2n - i + 1]$.

Example. The reversal of the word $ABBABAAB$ is $BAABABBA$.

Proposition 3.2. *Any word will have the same foldability as its reversal.*

Proof. The foldability of a word is determined by the order in which the letters are positioned. This can be easily seen because an B is between two As it will fold differently than if it is between two Bs. Note that it does not matter where in the word this letter is, because the starting letter can be changed without a change in foldability. So since when a word is reversed there is no change to the neighbors of each letter, the property that causes foldability is not altered, and the foldability does not change. Therefore, a word and its reversal will have the same foldability. □

Definition. A *transformation* on a word is the changing of the order of letters in a word without changing the foldability.

Definition. The *inverse* of a word is a word that has As where the original has Bs and vice versa.

Example. The inverse of the word $AABABBBABA$ is $BBABAAABAB$.

Proposition 3.3. *The inverse of a word is a transformation on that word.*

Proof. In order to show that the inverse of a word is a transformation, we need to show that taking the inverse does not effect the foldability.

So, we know that the foldability of a word is determined by the length of the word and the order of the letters and complements. However, because there is no real distinction between which is the letter and which is the complement. So as long as all of the letters are switched with all of the complements the order is not changed. Also since we are not adding or taking away letters or complements, the length is not changed either. Therefore by taking the inverse of a word, the foldability is not effected and the new word will have the same foldability. □

Definition. A *reversible pair* is an adjacent pair that can be reversed without changing the foldability of the word.

Example. Given the word $AAABB(AB)B$ the adjacent pair can be reversed to form the word $AAABB(BA)B$ without changing the foldability.

Proposition 3.4. *For every word made up of one letter and complement, given subwords Q and R (possibly empty words), $k(QAABAR) = k(QABAAR)$ if B can only bond with its neighbors, and $k(QBABBR) = k(QBBABR)$ if A can only bond with its neighbors.*

Proof. Let Q and R be subwords (possibly the empty word) of W and H. First let $W = QAABAR$ and $H = QABAAR$. We are given that for both W and H, the only possible bondings for B form adjacent pairs, because of that whatever bond B forms will not affect the rest of the word and we can think of it as being removed. So, take W, regardless of B bonds to the left or the right, removing that bond will leave us with $QAAR$. Now note that we receive the same result in H regardless of if B bonds to the left or the right. This means $k(W) = 2 \cdot k(QAAR) = k(H)$, or $k(W) = k(H)$.

Now let $W = QBABBR$ and $H = QBBABR$. Similar to above, we are given that in both W and H that A can only bond to its neighbors, and since this forms an adjacent pair we can think of it as being removed. Also similar to above, regardless of if A bonds to the left or the right in both W and H, by removing this bond we are left with $QBBR$. This means $k(W) = 2 \cdot k(QBBR) = k(H)$, or $k(w) = k(H)$.

Therefore, given subwords Q and R (possibly empty words), $k(QAABAR) = k(QABAAR)$ if B can only bond with its neighbors, and $k(QBABBR) = k(QBBABR)$ if A can only bond with its neighbors. □

From these operations, we were able to take the list of all the possible words of length 2, 4, 6, and 8 and summarize them with representative words.

Definition. A *representative word* is a word that is used to represent all of the words that can be related to it by an operation.

Our findings are summarized in the chart below.

Length	Number of Possible words	Representative Words
2	2	AB
4	6	AABB, ABAB
6	20	AAABBB, AABABB, ABABAB
8	70	AAAABBBB, AAABBABB, AABBAABB, ABABBABA, ABAABBAB, ABABABAB

Each of the words above have unique foldability and represent all of the possible ways for words of those lengths to fold. This can be used in the future to study the possible structures for words with unique foldabilities.

4 Conclusion and Future Work

In conclusion we found that certain words can be concatenated to form larger words with predictable foldability. We then used that to show that there exists a word of length $2n$, where n is not one less than a prime number, can fold $n+1$ times. We also found operations that relate words with the same foldability and summarized all the words of length 2, 4, 6, and 8 with their representative words.

Future work that can be done in this area would be to continue finding operations that can be performed on a word without changing the foldability of the word. One could also look at the structures of the representative words and look for patterns that correlate to foldability. Another question to look at is finding another way to build words of length $2n$ that fold $n+1$ ways in an attempt to fill in the holes with the prime numbers.

References

[1] Prime puzzles problem 43 catalan numbers. `http://www.primepuzzles.net/problems/prob_043.htm`. Accessed: 2018-7-3.

[2] B. Bjorkman, G. Cochran, W. Gao, L. Keough, R. Kirsch, M. Phillipson, D. Rorabaugh, H. Smith, and J. Wise. k-Foldability of Words. *ArXiv e-prints*, October 2017.

[3] Francis Black, Elizabeth Drellich, and Julianna Tymoczko. Valid plane trees: combinatorial models for RNA secondary structures with Watson-Crick base pairs. *SIAM J. Discrete Math.*, 31(4):2586–2602, 2017.