Analyzing shRNA Sequences with Multistrand

Edward Tang[1], Soni Singh[2]
[1]Henry M. Gunn High School, [2]independent

INTRODUCTION

Folded RNA has myriad applications, especially in the medical field. For example, the use of RNAi, or RNA interference, involves the use of folded siRNA—small interfering ribonucleic acids, a type of double stranded RNA—to target and destroy messenger RNA before it can be translated. shRNA, or short-hairpin RNA, is the most common secondary structure predicted, and it is named for its structure, which includes a double-stranded RNA stem with mismatches and unpaired sequences and a loop at the terminal end with unpaired RNA bases (Svoboda et. al, 2006).

The shRNA analyzed included the TRCN0000323374 and TRCN0000323375 sequences, which code for a protein subunit of the coatomer complex, helping the Golgi and associated vesicles move protein and lipids to the endoplasmic reticulum (Sigma Aldrich). shRNA is commonly used for long-term knockdown of a target gene through RNAi—RNA interference—a process where the RNA binds to Argonaute proteins which in turn cleave mRNA for a certain gene, reducing the gene's expression (Moore et. al, 2010) (University of Massachusetts, n.d.). These two sequences specifically target both the mouse and human. They were analyzed for their function in human research and for their similar structure: for the mRNA strands formed by each structure, the 5' and 3' strands are the same length, with the 5' strand being from base 5 - 25 and 3' being from base 32 - 52 (NCBI Probe).

Multistrand is a simulation that models the folding of RNA under different conditions. Specifically, it deals with cotranscriptional RNA, which is RNA that folds as it is being transcribed. It is hypothesized that this folding is the basis for what forms the RNA functional structures in the completed transcript (Lai et al., 2013). This RNA can later be expressed, but scientists have found use for various structures that can be formed through its cotranscriptional folding. Cotranscriptional RNA folding is dependent on many factors such as the RNA transcript, the entropy—or order of a system—and many more. These factors are modeled through markov chains—models that predict sequences of events based on probability—which predict, through aggregates of data, the most likely configurations of RNA secondary structure (Schaeffer, 2013).

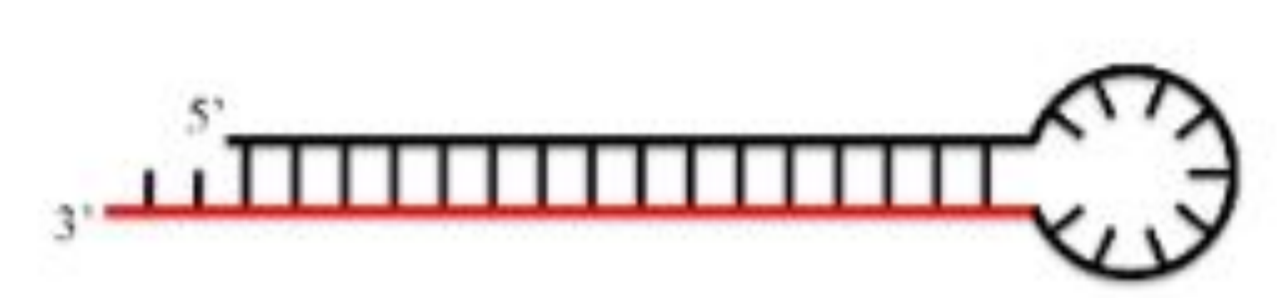


Figure 1: Example of a shRNA hairpin loop.

RESEARCH METHODOLOGIES

An experimental methodology was used to collect both quantitative and qualitative data regarding the formation of certain RNA shapes through the simulation Multistrand. The data was specifically drawn through the Hairpin trajectories tutorial, which modeled the secondary structure energy landscape of a certain shRNA sequence based on a Metropolis-biased random walk. (multistrand.org) A Metropolis-biased random walk is an algorithm based on Bayes' Theorem and Markov chains. Through this method, the probability for the transition from the initial, fully unpaired state to each other possible state is calculated, then simulated. From this second state, the probabilities to the next possible states are calculated and simulated until a user-set time limit is reached. Due to the project being conducted through the simulation, the assumptions made for the simulation are also present in the data; for example, it is assumed that the sequence of RNA is transcribed and does not make any pairing initially. The sequences used came from Sigma Aldrich and are used knockdown a gene that codes for human vesicle proteins.

The data provided by Multistrand highlights the structure and stability—in Gibbs free energy—of a shRNA sequence at a certain point in time after transcription. This data is based on state-change probabilities, and therefore, each run of the simulation may provide a new structure at each point in time. An example of the data at a certain point in time follows:

(((((((......))))))), t=0.000000052 seconds, dG= -8.58 kcal/mol

The parentheses represent paired bases while the periods represent unpaired ones. The time represents the time after transcription, and the stability is expressed in change in Gibbs free energy, where a lower free energy correlates with higher thermodynamic stability.

These data were analyzed to find similarities and differences both structurally and thermodynamically between the two similar sequences.

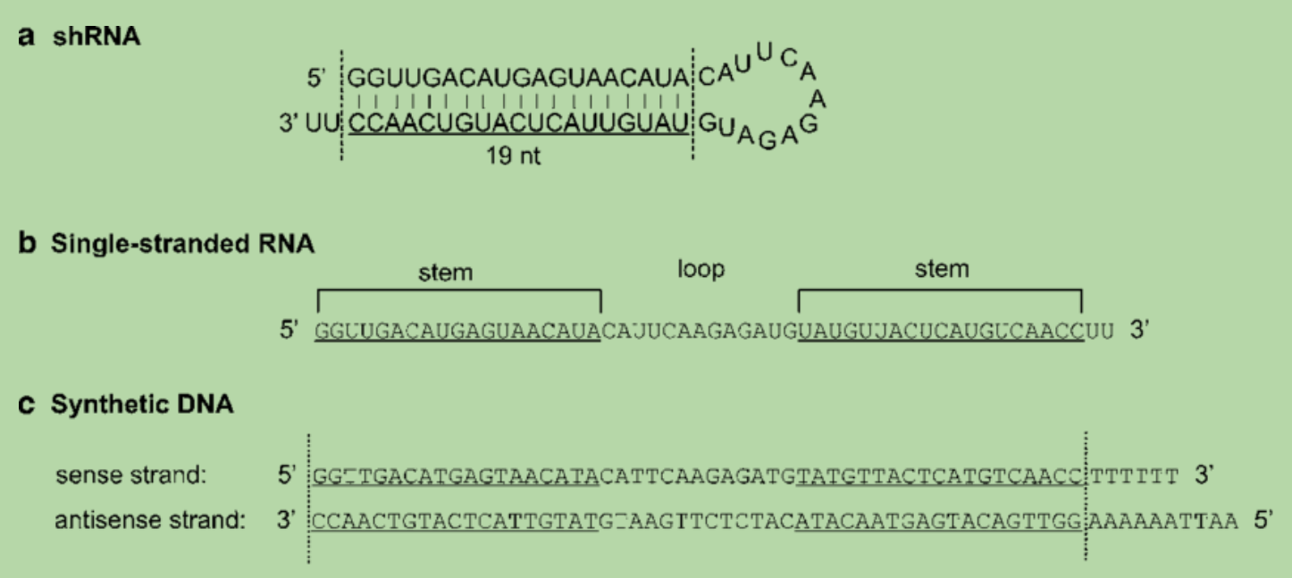
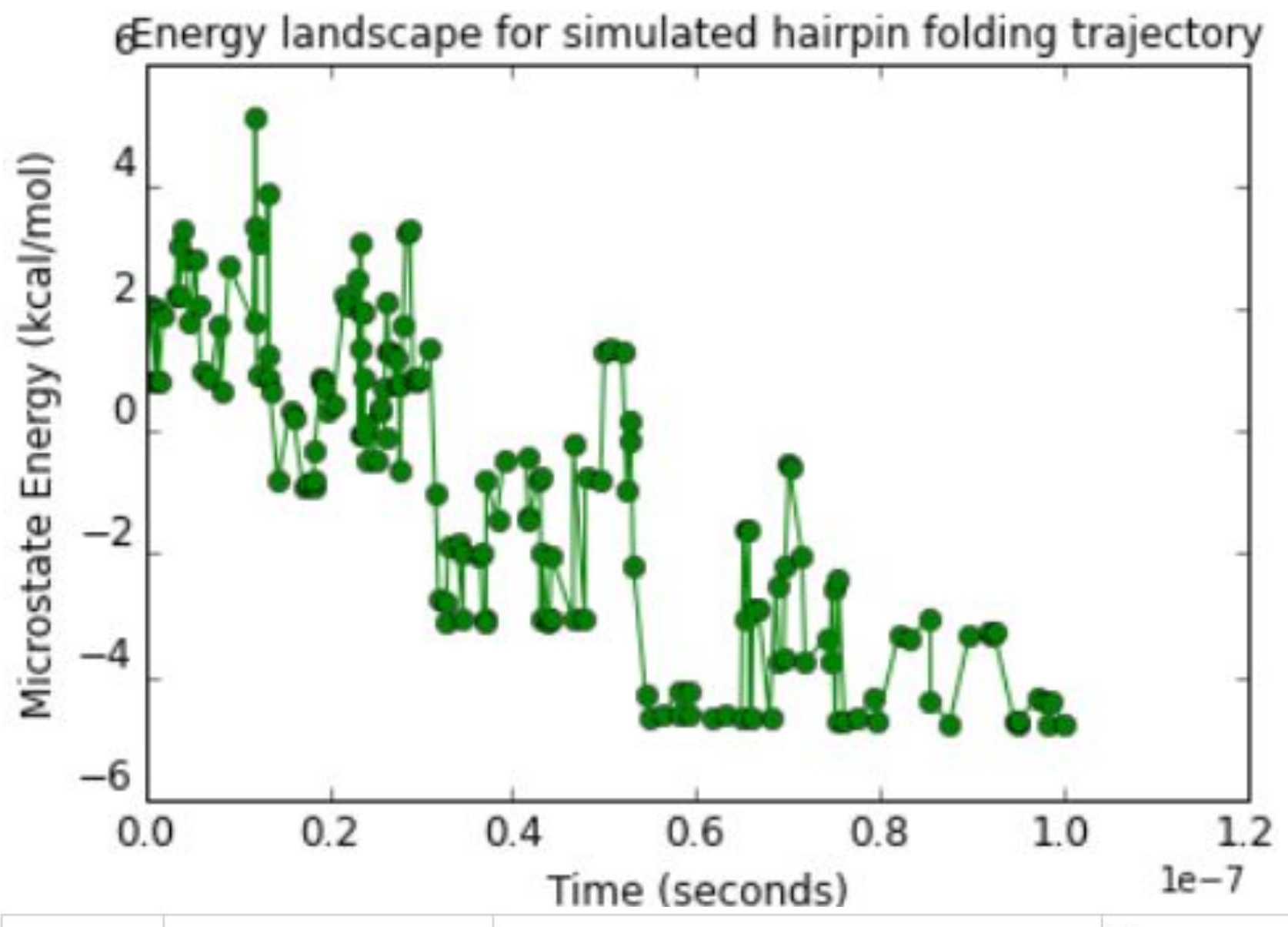

a shRNA
5' GGUUGACAUGAGUAACAUA CAUUCA
3' UU CCAACUGUACUCAUUGUAU GUAGAG A
19 nt
b Single-stranded RNA
stem
loop
stem
5' GGUUGACAUGAGUAACAUACAUUCAAGAGAUGUAUGUUACUCAUGUCAACCUU 3'
c Synthetic DNA
sense strand: 5' GGTTGACATGAGTAACATACATTCAAGAGATGTATGTTACTCATGTCAACCTTTTTT 3'
antisense strand: 3' CCAACTGTACTCATTGTATGTAAGTTCTCTACATACAATGAGTACAGTTGGAAAAAATTAA 5'

Figure 2: Comparison between RNA strand types

TRCN0000323374



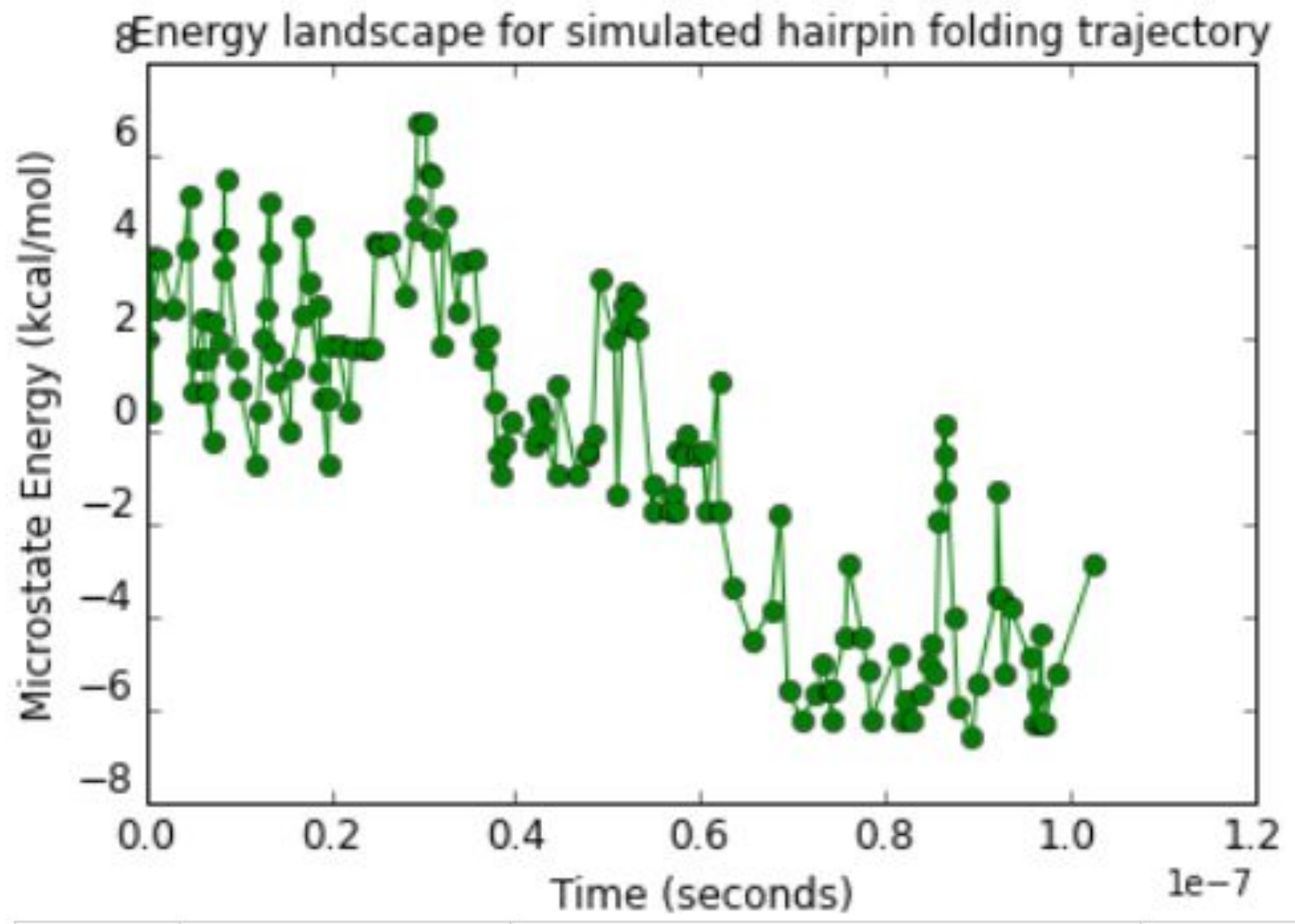
Frequency	Calculated Probability	Sequence	Times Appeared (E-8 s)	ΔG (kcal/mol)
1	1/1496	.....((((((((((((((((((....)))))))))))))))))))...	9.9	-31.14
1	1/1496	...(.(((((((((((((((((((....))))))))))))))))))))).	10.2	-30.1
2	2/1496	......((((((((((((((((((....))))))))))))))))))..... ...	9.9 (2)	-29.09

Mean	Minimum	Quartile 1	Median	Quartile 3	Maximum	Range	Standard Deviation
-2.253890374	-31.14	-3.16	-1.175	1.23	8.53	39.67	6.091549625

Frequency	Calculated Probability	Sequence	Times Appeared (E-8 s)	ΔG (kcal/mol)
37	37/1496	..((.....))..((((((((((((....))))))))))))((....)).......	2.8, 3, 4.2, 4.1, 4.8, 6.1, 7.2, 7.4, 8, 9.5, 9.5, 9.8, 10	-18.53

Structure Diagram
...............(.....)..................................
.....(......)..(.....)..................................
...............(.....)..................................
...(......)....(.....)..................................
...............(.....)..................................
....(.....)....(.....)..................................
....(.....)....(.....)............(.....)................
....(.....)....(.....)............(.....)...(....)........
....(.....)....(.....)............(.....)................
....(.....)....(.....)............(.....).....(....)......
....(.....)....(.....)...........((.....))....(....)......
....(.....)....(.....)...........((.....))................
....(.....)....(.....)...........((.....))....(....)......
....(.....).....................((.....))....(....)......
....(.....).....................((.....))................
....(.....)...................(..((.....)).)..............
....(.....)................(..(..((.....)).)..)...........
.........................(..(..((.....)).)..)..........
..(......)................(..(..((.....)).)..)..........
..(......)...................(..((.....)).)..............
..(......)..(......)..........(..((.....)).)..............
..(......)..(......).(.....)..(..((.....)).)..............
............(......).(.....)..(..((.....)).)..............
............(......)((.....)).(..((.....)).)..............
............(......)(((...))).(..((.....)).)..............
............(......)(((...))).(..((.....)).)....(...).....
...................(((...))).(..((.....)).)....(...).....
...................(((...)))....((.....))......(...).....
...................(((...)))....((.....)).....((...))....
....(.....).........(((...)))....((.....)).....((...))....
....(.....).........(((...)))....((.....)).....(.....)....
...................(((...)))....((.....)).....(.....)....
...................((.....))....((.....)).....(.....)....

TRCN0000323375



Frequency	Calculated Probability	Sequence	Times Appeared (E-8 s)	ΔG (kcal/mol)
14	14/1454	..........((((((((((((((((....)))))))))))))).......)))......	2.6, 3.7, 4.1, 4.2, 4.5, 6.1, 6.3, 7.2, 7.5, 8, 8.3, 8.5, 8.9, 9.3	-16.47
5	5/1454	.(....)...((((((((((((((((....)))))))))))))).......)))......	10.6, 7.2, 6.3, 4.9, 3.6	-15.81
8	8/1454	...........((((((((((((((....)))))))))))))).......))........	2.5, 3.8, 4.6, 4.7, 6, 8.2, 8.5, 8.7	-15.15

Mean	Minimum	Quartile 1	Median	Quartile 3	Maximum	Range	Standard Deviation
0.5527854195	-16.47	-1.3125	1.545	3.37	8.58	25.05	4.285292219

Frequency	Calculated Probability	Sequence	Times Appeared (E-8 s)	ΔG (kcal/mol)
14	14/1454	..........((((((((((((((((....)))))))))))))).......)))......	2.6, 3.7, 4.1, 4.2, 4.5, 6.1, 6.3, 7.2, 7.5, 8, 8.3, 8.5, 8.9, 9.3	-16.47

Structure Diagram
.........(((....)))...........(...................)......
.........((......))...........(...................)......
.........(((....)))...........(...................)......
.........(((....)))...........(.....(.....).......)......
.........((......))...........(.....(.....).......)......
.........((......))..................(.....).............
.(....)..((......))..................(.....).............
.(....)..((......))..........(....)..(.....).............
.(....)..((......))..................(.....).............
.........((......))..................(.....).............
.........(((....)))..................(.....).............
.........(((....)))..(....)..........(.....).............
.........(((....)))..(....)..............................
.........(((....)))..(....)...(...............)..........
.(....)..(((....)))..(....)...(...............)..........
.(....)..(((....)))..(....).(.(...............)..)........
.(....)..(((....)))..(....)...(...............)..........
.(....)..(((....)))..........(...............)..........
.........(((....)))..........(...............)..........
.........(((....)))......................................
.........((......))......................................
.........((......))..........(..........)................
.(....)..((......))..........(..........)................
.(....)..((......))..........((.......).)................
.(....)..((......))...........(.......)..................
.(....)..((......))..........((.......))................
.........((......))..........((.......))................
.........(((....)))..........((.......))................
.........(((....))).........(.((.......)))...............
.........(((....))).........(.(.........))...............
.........(((....))).........(.(.........))........(.....).
.........(((....))).........(...........)........(.....).

CONCLUSIONS AND ANALYSIS

The data shows that TRCN0000323374 has a higher maximum stability when compared to TRCN0000323375, shown by the much lower Gibbs free energy values present in the data set for TRCN0000323375. Furthermore, it seemed that TRCN0000323375 was more variable. Even though it had the most stable structures, the top three structures in terms of stability were uncommon, appearing once, once, and twice respectively. Less stable structures—still more stable than TRCN0000323374's most stable structures—were more common and appeared with higher frequency. On the other hand, the most stable structures for TRCN0000323374 appeared much more often, but had around twice as much Gibbs free energy, showing that it was more unstable.

It is difficult to determine which sequence is more viable for manufacturing, as TRCN0000323374 has a higher maximum stability, but TRCN0000323375's most stable state is its most common state. That being said, TRCN0000323374's average Gibbs free energy was much lower than that of TRCN0000323375—almost twice as low—showing that TRCN0000323374 is much more stable. TRCN0000323374, however, also had a higher standard deviation and range, suggesting that it has much more variation in terms of stability than TRCN0000323375. Finally, the most frequent structure for TRCN0000323374 was much more common but only slightly more stable than that of TRCN0000323375—by around 2 kj/mol.

It seems that, for the majority of applications where stable RNA is necessary to building devices, TRCN0000323374 is a better option due to its general stability. That being said, it is not applicable in all uses, due to its higher variability in terms of stability and the dependency of certain structures and shapes for specific applications, as in RNA genetic knockdown.

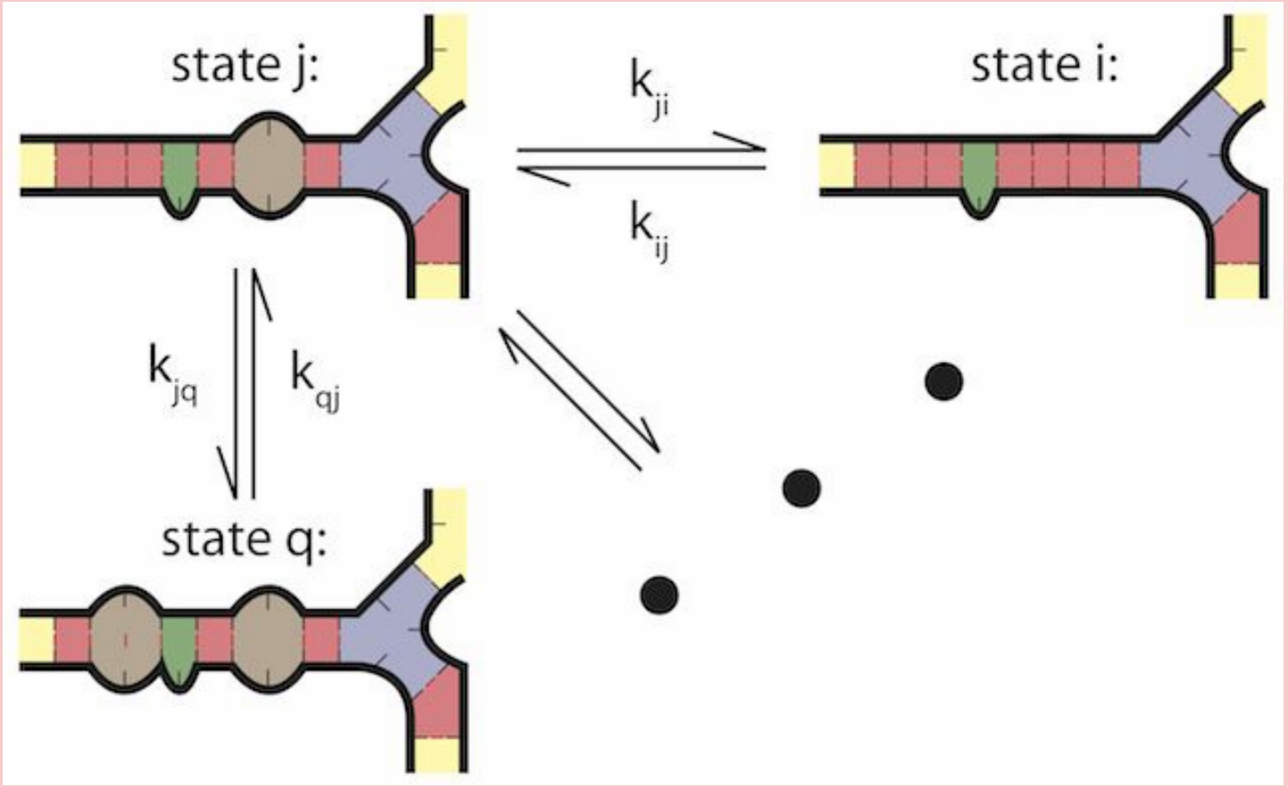


Figure 3: Demonstration of different structure states

ACKNOWLEDGEMENTS / REFERENCES

Special thanks to Angela Merchant and Soni Singh for helping make this project possible.

Works Cited:

On back of poster.

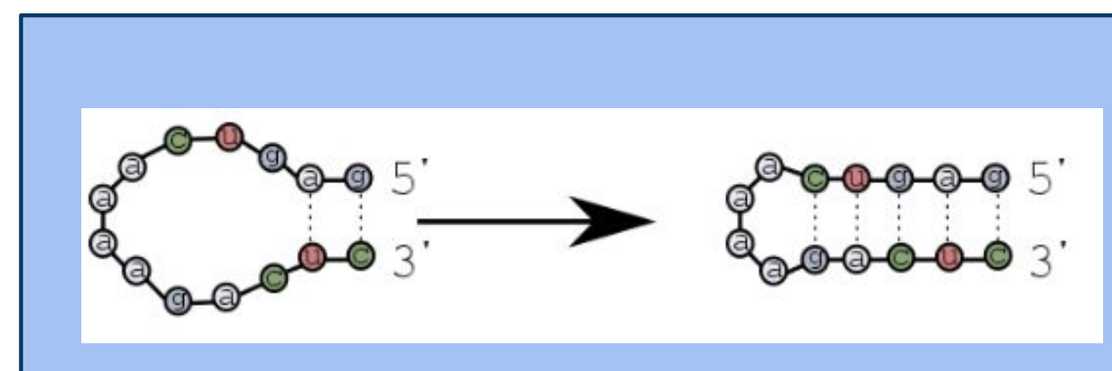


Figure 4: Hairpin loop folding towards inside

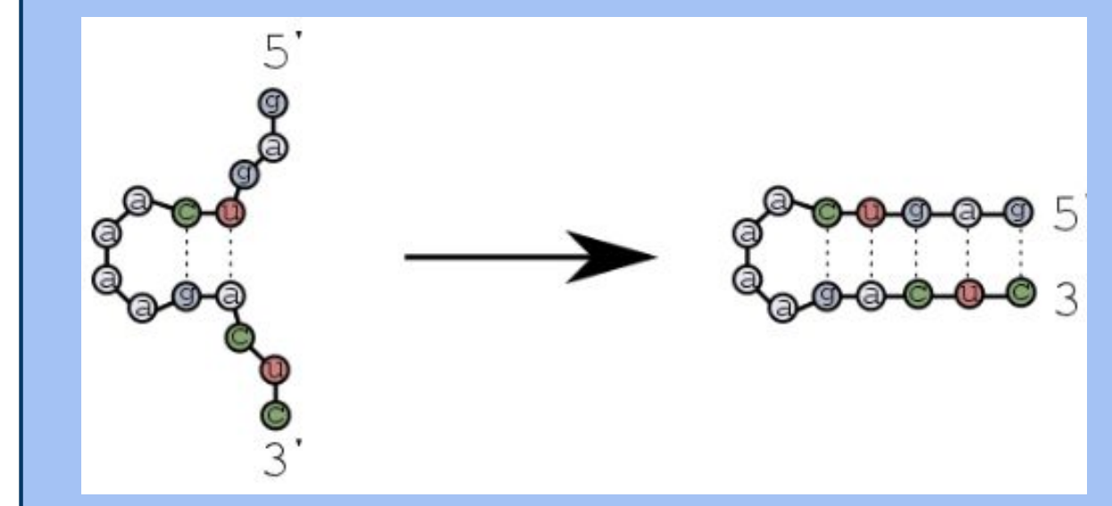


Figure 5: Hairpin loop folding towards outside

IMPLICATIONS / NEXT STEPS

The significant difference between the stability and folding structures of the two sequences analyzed shows that while they are similar sizes in terms of base pairs and have the same function of coding for the same protein, their applications in terms of manufacturing devices and functions outside of their designed functions vary. The data can be used to determine which sequences are better for each application, as the most common and most stable structures are a factor in testing whether a sequence fits a certain function.

In terms of future steps, there can be two phases: first, these two sequences can be analyzed further. The first step would be to continue to run trials for the same simulation to make sure that the data is accurate. Since I had limited processing power and time for data analysis, I was only able to conduct 10 trials per sequence. While this yielded close to 1500 data points for each of the sequences, 10 trials is still too few trials to draw accurate data, as there would only be around 40 to 50 different structures at max for each time point. Next, the same sequences could be analyzed for different types of results, such as how long it takes for it to reach a stable form without changing. This test is also present in the Multistrand simulation along with a dozen different tests—though some are not applicable to the analyzed sequences. Testing for this other data could draw a more comprehensive picture of the sequences and would give researchers more information that could be used to select the optimal sequence to use, along with experimental details such as the optimal time to allow an RNA sequence to fold. The second phase of future steps would be to test other RNA sequences. Since there are myriad sequences in existence, testing many more sequences would allow researchers to differing needs. Therefore, they would not need to settle on a sequence that has already been researched but is not optimal for a given task; instead, they could select the perfect sequence for each application. If the aforementioned tests could be performed on many sequences and the data compiled into a database, manufacturing with RNA would become much easier and more efficient.