Inferring RNA Ancestors

COMP 598: Advanced Computation Biology Methods & Research

Project Report

Haji Mohammad Saleem
260508983

Introduction

Multiple species that inhabit this planet have been shown to share genetic similarities. Phylogenetics is the study of evolutionary history and relationships among such individuals or groups of organisms. Phylogenies are branching diagrams that represent the ancestral relationships among a set of species. Internal nodes in a phylogeny tree are referred to as ancestors. When an ancestor diverges, the two species are assumed to evolve independently of each other.

One of the ways to construct a phylogenetic tree of related set of genetic sequences is through maximum parsimony, a non-parametric statistical model, first introduced by Walter M. Fitch in 1971. A phylogeny constructed through parsimony prefers evolution with least changes.

This project implements Sankoff's algorithm [1,2], a generalized version of the Fitch's algorithm. We use RNA families from the Rfam database for the analysis [3], extracted in the Stockholm format [4]. In this project with compute the sequence and predict the secondary structure of each ancestor and analyze its conservation in relation to the consensus structure.

Methods

We extract the sequences of the leaf nodes and the consensus structure of the RNA family from the Stockholm files obtained from the RFAM database. We also acquire the phylogeny tree for the family from the database. Next, based on the phylogenetic tree, we predict the ancestral sequence, using the Sankoff Algorithm. The algorithm uses the following cost matrix to account for gaps:

	A	C	G	U	-
A	0	2	1	2	2
C	2	0	2	1	2
G	1	2	0	2	2
U	2	1	2	0	2
-	2	2	2	2	0

The weights for leaf nodes in the phylogeny is initialized such that the present nucleotide has a weight 0 while the other possible nucleotides have a weight "inf". Next for each ancestral sequence, we calculate the index weight as

$$S_a(i) = min_j\left[c_{ij} + S_l(j)\right] + min_k[c_{ik} + S_r(k)]$$

where $S_a(i)$ is the smallest cost for node a for the state i. The states are chosen to minimize the cost to move to state j and k for left and right child. Thus, for each index of the ancestral sequence, we choose the nucleotide with the minimum cost.

To conserve the base pair dependencies of the consensus secondary structure, we remove the possibility of gaps at positions with base pairs in the consensus structure and use a matching nucleotide to complete base pairs at base pair closing indices.

Data

For our initial analysis (Objective 1-7), we use the **RF00754:** microRNA mir-279 RNA family

Code

```
python sankoff_algo.py -h
usage: sankoff_algo.py [-h] -r RFAM [-p] [-e]

Implementation of the Sankoff Algorithm

optional arguments:
  -h, --help            show this help message and exit
  -r RFAM, --rfam RFAM  RNA family id
  -p, --printit         Print additional information
  -e, --extended        Use Sankoff with base pair preservation
```

Results

Stockholm Parser (Objective 1)

We implement the `Stockholm` class in `stockolm_parser.py`. It extracts the consensus structure and the sequence of all the member species.

Output

```
RNA family RF00754
Consensus Secondary Structure
                                 ::::<<<<<<<<<<<<<<<<<<<<<<.<<<<<<----<<<<<__________...____>>>>>>>>>>.>>>>>>>>>>>>>>>>>>>...>>::
Sequence Alignment
AADE01000363.1/2006-2100         UCAUACUACUGUUUUUAGUGAGUGAGG-GUCCAGUGUUUCACAUUGAUUUUCUUA-GUAUUUGUGACUAGAU-CCACACUCAUUAAUAACGGUA---GUUC
AANI01014375.1/46325-46420       GCGUACUACUGUUUUUAGUGAGUGAGG-GUCCAGUGUUUCACAUUCGUUUUUUCAAGUAUUUGUGACUAGAU-CCACACUCAUUAACAACGGUA---GUUC
AAGE02017556.1/117017-116925     CAUCCCAACGGUUGUCGAUGGGUGUGA-AUCUAGUGUUUCACAUGAUUUUCGAUA-GCC--UGUGACUAGAU-CCACACUCAUUAACAAAAGUU---GCCG
AAPT01018378.1/277640-277732     CCGUAUUACUGUUUUUAGUGGGUGAGG-GUCCAGUGUUUCACAUUGUUUAUUUG---UAUUUGUGACUAGAU-CCACACUCAUUAAAAACGGUA---GUUC
AAJJ01000015.1/255293-255381     AAUUUGAUCCGUUCUUGAUGGGUUCGG-GUCUAGUGG--CACGGUUUUUUCA-----ACUUCGUGACUAGAU-CCACACUCAUUAAGGAAGUUU---CACA
AEKZ01010226.1/4670-4585         UUUCCUGAAUUUGCCAAAUGAGUGAAG-GUCUAGUG---CACAGAAAAUGAA-------AUUGUGACUAGAU-CCACACUCAUUAAGUACGUUC---AGGU
JN211060.1/1022-1116             UCAUACUACUGUUUUUAGUGGGUGGGG-GUCCAGUGUUUCACAUUGAUUUUCUUA-GUAUUUGUGACUAGAU-CCACACUCAUUAAUAACGGUA---GUUC
AAJJ01003219.1/11291-11201       UGGAGCUCUCGGUGUAAGCCAGUGUUCAGUCUAUUGUUUCACAUUGGUUUCG-------AUUGUGACUAGAUCGAACACUCGCUUGCAACCUGG---GUUU
AAZX01001678.1/20489-20404       CCAGCCGAUUGUACUGAGUGAGUGAUG-GUCUGGUG-----CACGGUUUAUC-----GAUCUGUGACUAGAU-CCACACUCAUUAAGUACGUUC---GGCU
AAPP01019575.1/159482-159391     UACUACUACUGUUUUUAGUGGGUGAGG-GUCCAGUGUUUCACAUUGAUUUCUG----UAUUUGUGACUAGAU-CCACACUCAUUAAUAACGGUA---GUUC
AADG06007490.1/6579-6668         UUGUUCGAUGGCCUUGGAUGGGUUUGA-AUUCAG---UCCACGUU--UUUUAUUU-UUAUUCGUGACUAGAU-CCACACUCAUCCAAGGAAAUC---GAGC
AAWU01001985.1/8444-8349         GCUUCCCACUAUUGUCGAUGGGUGUGA-AUCUAGUGGUUCACAUGAGCUUUGCC--A-AACUGUGACUAGAU-CCACACUCAUUAACAAGAGUGCUCGGAA
```

Sankoff Algorithm (Objective 2-5)

We implement the `Sankoff` class in `sankoff_algo.py`. The initial cost matrix is designed for the four nucleotides. The class uses the RNA family id along with the `Stockholm` object. The phylogeny tree file is read by `readTree()`. To get the ancestral sequences, first use `getAncestorWeight()` and then use `getAncestorSeq()`.To include gaps, the cost matrix is upgraded to account for gap transitions, executed by `includeGaps()`. We use wrappers for `RNAfold` and `RNAdistance`.

```
python sankoff_algo.py -r RF00754 -p
```

Output

```
                                    Sankoff Analysis for RF00754 Family.
-------------------------------------------------------------------------------------------------
                    RF00754 Family Tree
-------------------------------------------------------------------------------------------------
           ________ AAPP01019575.1/159482-159391
  ________|
 |        |         _________ AADE01000363.1/2006-2100
 |        |________|
 |                 |_________ JN211060.1/1022-1116
 |
 |                  _________ AAGE02017556.1/117017-116925
 |         ________|
_|        |        |_________ AAWU01001985.1/8444-8349
 |________|
 |        |         _________ AAJJ01000015.1/255293-255381
 |        |________|
 |                 |_________ AADG06007490.1/6579-6668
 |
 |         ________ AAPT01018378.1/277640-277732
 |        |
 |________|         _________ AANI01014375.1/46325-46420
          |        |
          |________|                    _________ AAZX01001678.1/20489-20404
                   |          ________|
                   |_________|        |_________ AEKZ01010226.1/4670-4585
                             |
                             |________ AAJJ01003219.1/11291-11201

-------------------------------------------------------------------------------------------------
                    Leaf Nodes: 12, Ancestor Nodes: 11, Sequence Length: 101
-------------------------------------------------------------------------------------------------
                                     Cost Matrix
-------------------------------------------------------------------------------------------------
                             A        C        G        U        -
                    A        0        2        1        2        2
                    C        2        0        2        1        2
                    G        1        2        0        2        2
                    U        2        1        2        0        2
                    -        2        2        2        2        0
-------------------------------------------------------------------------------------------------
                                     Ancestors
-------------------------------------------------------------------------------------------------
Ancestor                              Left Child                            Right Child
1                                     AADE01000363.1/2006-2100              JN211060.1/1022-1116
2                                     AAPP01019575.1/159482-159391          1
3                                     AAGE02017556.1/117017-116925          AAWU01001985.1/8444-8349
4                                     AAJJ01000015.1/255293-255381          AADG06007490.1/6579-6668
5                                     3                                     4
6                                     AAZX01001678.1/20489-20404            AEKZ01010226.1/4670-4585
7                                     6                                     AAJJ01003219.1/11291-11201
8                                     AANI01014375.1/46325-46420            7
9                                     AAPT01018378.1/277640-277732          8
10                                    2                                     5
11                                    10                                    9
-------------------------------------------------------------------------------------------------
```

```
                                        Leaf Nodes
----------------------------------------------------------------------------------------------------
Accession                      Sequence
AADE01000363.1/2006-2100       UCAUACUACUGUUUUUAGUGAGUGAGG-GUCCAGUGUUUCACAUUGAUUUUCUUA-GUAUUUGUGACUAGAU-CCACACUCAUUAAUAACGGUA---GUUC
AADG06007490.1/6579-6668       UUGUUCGAUGGCCUUGGAUGGGUUUGA-AUUCAG---UCCACGUU--UUUUAUUU-UUAUUCGUGACUAGAU-CCACACUCAUCCAAGGAAAUC---GAGC
AAGE02017556.1/117017-116925   CAUCCCAACGGUUGUCGAUGGGUGUGA-AUCUAGUGUUUCACAUGAUUUUCGAUA-GCC--UGUGACUAGAU-CCACACUCAUUAACAAAAGUU---GCCG
AAJJ01000015.1/255293-255381   AAUUUGAUCCGUUCUUGAUGGGUUCGG-GUCUAGUGG--CACGGUUUUUUCA-----ACUUCGUGACUAGAU-CCACACUCAUUAAGGAAGUUU---CACA
AAJJ01003219.1/11291-11201     UGGAGCUCUCGGUGUAAGCCAGUGUUCAGUCUAUUGUUUCACAUUGGUUUCG-------AUUGUGACUAGAUCGAACACUCGCUUGCAACCUGG---GUUU
AANI01014375.1/46325-46420     GCGUACUACUGUUUUUAGUGAGUGAGG-GUCCAGUGUUUCACAUUCGUUUUUCAAGUAUUUGUGACUAGAU-CCACACUCAUUAACAACGGUA---GUUC
AAPP01019575.1/159482-159391   UACUACUACUGUUUUUAGUGGGUGAGG-GUCCAGUGUUUCACAUUGAUUUCUG----UAUUUGUGACUAGAU-CCACACUCAUUAAUAACGGUA---GUUC
AAPT01018378.1/277640-277732   CCGUAUUACUGUUUUUAGUGGGUGAGG-GUCCAGUGUUUCACAUUGUUUAUUUG---UAUUUGUGACUAGAU-CCACACUCAUUAAAAACGGUA---GUUC
AAWU01001985.1/8444-8349       GCUUCCCACUAUUGUCGAUGGGUGUGA-AUCUAGUGGUUCACAUGAGCUUUGCC--A-AACUGUGACUAGAU-CCACACUCAUUAACAAGAGUGCUCGGAA
AAZX01001678.1/20489-20404     CCAGCCGAUUGUACUGAGUGAGUGAUG-GUCUGGUG-----CACGGUUUAUC-----GAUCUGUGACUAGAU-CCACACUCAUUAAGUACGUUC---GGCU
AEKZ01010226.1/4670-4585       UUUCCUGAAUUUGCCAAAUGAGUGAAG-GUCUAGUG---CACAGAAAAUGAA-------AUUGUGACUAGAU-CCACACUCAUUAAGUACGUUC---AGGU
JN211060.1/1022-1116           UCAUACUACUGUUUUUAGUGGGUGGGG-GUCCAGUGUUUCACAUUGAUUUUCUUA-GUAUUUGUGACUAGAU-CCACACUCAUUAAUAACGGUA---GUUC
----------------------------------------------------------------------------------------------------
                                        Ancestor Nodes
----------------------------------------------------------------------------------------------------
Accession                      Sequence
1                              UCAUACUACUGUUUUUAGUGAGUGAGG-GUCCAGUGUUUCACAUUGAUUUUCUUA-GUAUUUGUGACUAGAU-CCACACUCAUUAAUAACGGUA---GUUC
2                              UAAUACUACUGUUUUUAGUGGGUGAGG-GUCCAGUGUUUCACAUUGAUUUCCGUA-GUAUUUGUGACUAGAU-CCACACUCAUUAAUAACGGUA---GUUC
3                              CAUCCCAACGAUUGUCGAUGGGUGUGA-AUCUAGUGGUUCACAUGAGCUUCGACA-ACAACUGUGACUAGAU-CCACACUCAUUAACAAAAGUGCUCGCAA
4                              AAGUUCAACCGCCCUGGAUGGGUUCGA-AUCCAGUGGUCCACGGUUUUUCAUUU-UAAUUCGUGACUAGAU-CCACACUCAUCAAAGAAAAUC---CACA
5                              AAUUCCAACGGUUCUCGAUGGGUGUGA-AUCUAGUGGUCCACAUGAUUUUCACU--ACAACCGUGACUAGAU-CCACACUCAUUAAAAAAAAAUU---GACA
6                              CCACCCGAAUGUACCAAAUGAGUGAAG-GUCUAGUG---CACACAAAAUAAA-----GAACUGUGACUAGAU-CCACACUCAUUAAGUACGUUC---AGCU
7                              UCAACCGAUCGGACUAAGCCAGUGAUCAGUCUAGUGUUUCACACAGAUUACA-------AUUGUGACUAGAUCCAACACUCACUAACAACCUGC---GGCU
8                              GCGAACUACUGUUCUAAGUGAGUGAGG-GUCCAGUGUUUCACAUUCGUUUUAUCAAGUAAUUGUGACUAGAU-CCACACUCAUUAACAACGGUA---GUUC
9                              CCGUACUACUGUUUUUAGUGAGUGAGG-GUCCAGUGUUUCACAUUGGUUAUUUC---UAUUUGUGACUAGAU-CCACACUCAUUAAAAACGGUA---GUUC
10                             UACUACAACUGUUUUUAAUGGGUGAGA-AUCCAGUGGUUCACAUUAAUUUCAUU--GUAUUUGUGACUAGAU-CCACACUCAUUAACAAAAGUA---GACA
11                             UCAUACUACUGUUUUUAGUGGGUGAGG-GUCCAGUGUUUCACAUUGUUUUUUUC---UAUUUGUGACUAGAU-CCACACUCAUUAACAACGGUA---GUUC
----------------------------------------------------------------------------------------------------
                                        Secondary Structures
----------------------------------------------------------------------------------------------------
Accession                      Sequence
1                              UCAUACUACUGUUUUUAGUGAGUGAGG-GUCCAGUGUUUCACAUUGAUUUUCUUA-GUAUUUGUGACUAGAU-CCACACUCAUUAAUAACGGUA---GUUC
Dist: 20                       ......(((((((.((((((((((.((.(((.((....((((((..................))))))).))).)).)))))))))).)))))))).......
2                              UAAUACUACUGUUUUUAGUGGGUGAGG-GUCCAGUGUUUCACAUUGAUUUCCGUA-GUAUUUGUGACUAGAU-CCACACUCAUUAAUAACGGUA---GUUC
Dist: 20                       ......(((((((.((((((((((.((.(((.((....(((((..................))))))).))).)).)))))))))).))))))).......
3                              CAUCCCAACGAUUGUCGAUGGGUGUGA-AUCUAGUGGUUCACAUGAGCUUCGACA-ACAACUGUGACUAGAU-CCACACUCAUUAACAAAAGUGCUCGCAA
Dist: 58                       .......((..((((.((((((((((..((((((((((((....)))))...(((......))).)))))))..)))))))))).))))..))........
4                              AAGUUCAACCGCCCUGGAUGGGUUCGA-AUCCAGUGGUCCACGGUUUUUCAUUU-UAAUUCGUGACUAGAU-CCACACUCAUCAAAGAAAAUC---CACA
Dist: 68                       .............((.(((((((.....(((...(((((.((((.((..........)).))))))))))))).....)))))))..)).............
5                              AAUUCCAACGGUUCUCGAUGGGUGUGA-AUCUAGUGGUCCACAUGAUUUUCACU--ACAACCGUGACUAGAU-CCACACUCAUUAAAAAAAAAUU---GACA
Dist: 70                       ................((((((((((..((((((((((.....(((...))).......)))...)))))))..)))))))))).................
6                              CCACCCGAAUGUACCAAAUGAGUGAAG-GUCUAGUG---CACACAAAAUAAA-----GAACUGUGACUAGAU-CCACACUCAUUAAGUACGUUC---AGCU
Dist: 28                       ......((((((((..((((((((..(.(((((((....((((..................))))))))))).)..))))))))..)))))))).......
7                              UCAACCGAUCGGACUAAGCCAGUGAUCAGUCUAGUGUUUCACACAGAUUACA-------AUUGUGACUAGAUCCAACACUCACUAACAACCUGC---GGCU
Dist: 70                       ....((....))....((((((((....((((((....(((((..................)))))))))))....)))).................))))
8                              GCGAACUACUGUUCUAAGUGAGUGAGG-GUCCAGUGUUUCACAUUCGUUUUAUCAAGUAAUUGUGACUAGAU-CCACACUCAUUAACAACGGUA---GUUC
Dist: 26                       ..(((((((((((...((((((((.((.(((.((....(((((..................))))))).))).)).))))))))...)))))))...))))
9                              CCGUACUACUGUUUUUAGUGAGUGAGG-GUCCAGUGUUUCACAUUGGUUAUUUC---UAUUUGUGACUAGAU-CCACACUCAUUAAAAACGGUA---GUUC
Dist: 16                       ......((((((((((((((((((.((.(((.((....(((((..................))))))).))).)).)))))))))))))))))).......
10                             UACUACAACUGUUUUUAAUGGGUGAGA-AUCCAGUGGUUCACAUUAAUUUCAUU--GUAUUUGUGACUAGAU-CCACACUCAUUAACAAAAGUA---GACA
Dist: 44                       .......(((....((((((((((.(..(((...((((.((((..................)))))))))))..).))))))))))....)))........
11                             UCAUACUACUGUUUUUAGUGGGUGAGG-GUCCAGUGUUUCACAUUGUUUUUUUC---UAUUUGUGACUAGAU-CCACACUCAUUAACAACGGUA---GUUC
Dist: 20                       ......(((((((.((((((((((.((.(((.((....(((((..................))))))).))).)).)))))))))).))))))).......
```

Base pair conservation (Objective 6,7)

To preserve the base pairs, we propose two things:

1) Remove the possibility of gaps at base pair sites.
2) Select a matching nucleotide at the closing base pair site to increase the likelihood of base pair formation.

```
python sankoff_algo.py -r RF00754 -p -e
```

Output

```
                                    Sankoff Analysis for RF00754 Family with base pair conservation.
-------------------------------------------------------------------------------------------------
                                    Secondary Structures
-------------------------------------------------------------------------------------------------
Accession                           Sequence
1                                   UCAUACUACUGUUUUUAGUGAGUGAGG-GUCCAGUGUUUCACAUUGAUUUUCUUA-GUAUUUGUGACUGGAU-CCUCACUCAUUAAAAACGGUA---GUUC
Dist: 8                             ......((((((((((((((((((((.((((((....(((((.................)))))))))).)))))))))))))))))))).......
2                                   UAAUACUACUGUUUUUAGUGGGUGAGG-GUCCAGUGUUUCACAUUGAUUUCCGUA-GUAUUUGUGACUGGAU-CCUCACUCAUUAAAAACGGUA---GUUC
Dist: 8                             ......((((((((((((((((((((.((((((....(((((.................)))))))))).)))))))))))))))))))).......
3                                   CAUCCCAACGAUUGUCGAUGGGUGUGA-AUCUAGUGGUUCACAUGAGCUUCGACA-ACAACUGUGACUAGAU-UCACACUCAUUGACAAUCGUUCUCGGAA
Dist: 38                            ..(((.((((((((((((((((((((.((((((((((((....)))))...(((......))).)))))))).))))))))))))))))))...))).
4                                   AAGUUCAACCGCCCUGGAUGGGUUCGA-AUCCAGUGGUCCACGGUUUUUUCAUUU-UAAUUCGUGGCUGGAU-UCGAACUCAUCCAGGGCGGUU---GACA
Dist: 18                            ..(((.((((((((((((((((((((.((((((....((((((.((..........)).)))))))))))).))))))))))))))))))...))).
5                                   AAUUCCAACGGUUCUCGAUGGGUGUGA-AUCUAGUGGUCCACAUGAUUUUCACU--ACAACUGUGGCUAGAU-UCACACUCAUUGAGAACCGUU---GGCA
Dist: 8                             ......((((((((((((((((((((.((((((....(((((.................)))))))))).)))))))))))))))))))).......
6                                   CCACCCGAAUGUACCAAAUGAGUGAAG-GUCUAGUG--ACACACAAAAUAAA-----GAACUGUGUCUAGAU-CUUCACUCAUUUGGUACGUUC---GGCU
Dist: 2                             ...((.((((((((((((((((((((.((((((....(((((.................)))))))))).)))))))))))))))))))...)).. 
7                                   UCAACCGAUCGGACUAAGCCAGUGAUCAGUCUAGUGUUUCACACAGAUUACA-------AUUGUGACUAGAUCGAUCACUGGCUUAGUCCGAUC---GGCU
Dist: 0                             ....((((((((((((((((((((.((((((....(((((.................)))))))))).)))))))))))))))))))...)).. 
8                                   GCGAACUACUGUUCUAAGUGAGUGAGG-GUCCAGUGUUUCACAUUCGUUUUAUCAAGUAAUUGUGACUGGAU-CCUCACUCAUUUAGAACGGUA---GUUC
Dist: 8                             ..((((((((((((((((((((.((((((....(((((.................)))))))))).)))))))))))))))))))...))))
9                                   CCGUACUACUGUUUUUAGUGAGUGAGG-GUCCAGUGUUUCACAUUGGUUAUUUC---UAUUUGUGACUGGAU-CCUCACUCAUUAAAAACGGUA---GUUC
Dist: 8                             ......((((((((((((((((((((.((((((....(((((.................)))))))))).)))))))))))))))))))).......
10                                  UACUACAACUGUUUUUAAUGGGUGAGA-AUCCAGUGGUUCACAUUAAUUUCAUU--GUAUUUGUGACUGGAU-UCUCACUCAUUAAAAACAGUU---GUCA
Dist: 8                             ......((((((((((((((((((((.((((((....(((((.................)))))))))).)))))))))))))))))))).......
11                                  UCAUACUACUGUUUUUAGUGGGUGAGG-GUCCAGUGUUUCACAUUGUUUUUUUC---UAUUUGUGACUGGAU-CCUCACUCAUUAAAAACGGUA---GUUC
Dist: 8                             ......((((((((((((((((((((.((((((....(((((.................)))))))))).)))))))))))))))))))).......
```

Family Analysis (Objective 8)

We run the extended Sankoff algorithm on a set of RNA families.

RNA family	# Ancestors	Seq Length	RNA Dist	GC content	MFE freq
RF00434	16	130	606	877	0.5836985
RF00489	9	49	108	231	2.768614
RF00754	11	101	114	428	2.2431504
RF00951	23	52	366	329	9.90876
RF01313	4	57	86	100	0.193195
RF01318	11	37	64	141	4.966388
RF01320	9	37	24	110	5.537081
RF01696	14	70	0	511	11.298948
RF01909	24	155	3732	1723	0.2666189
RF01978	9	103	1076	310	0.41786508
RF02045	17	150	2304	893	0.160442971
RF02114	9	123	980	463	0.1883405
RF02353	11	84	260	561	3.522485
RF02375	12	233	1208	756	0.144271489
RF02506	13	95	244	794	4.4512913

Table 1: *Number of ancestors, length of sequence, RNA distance, GC content and MFE frequency of various RNA families*

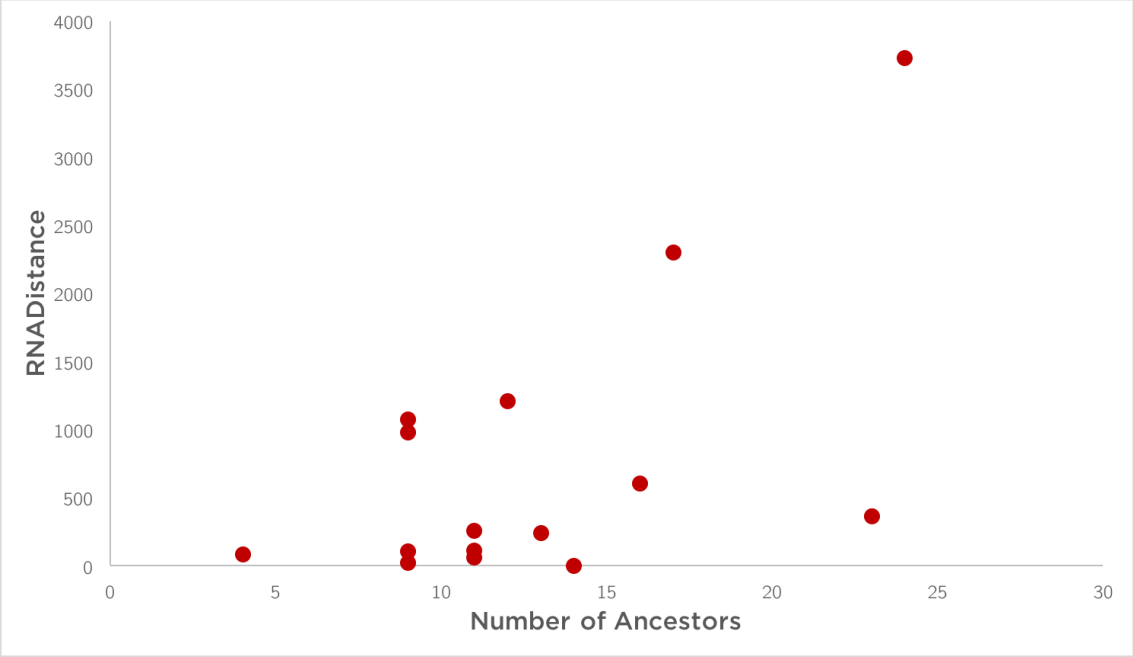

Figure 1: *Number of ancestors vs the net RNADistance in an RNA family*

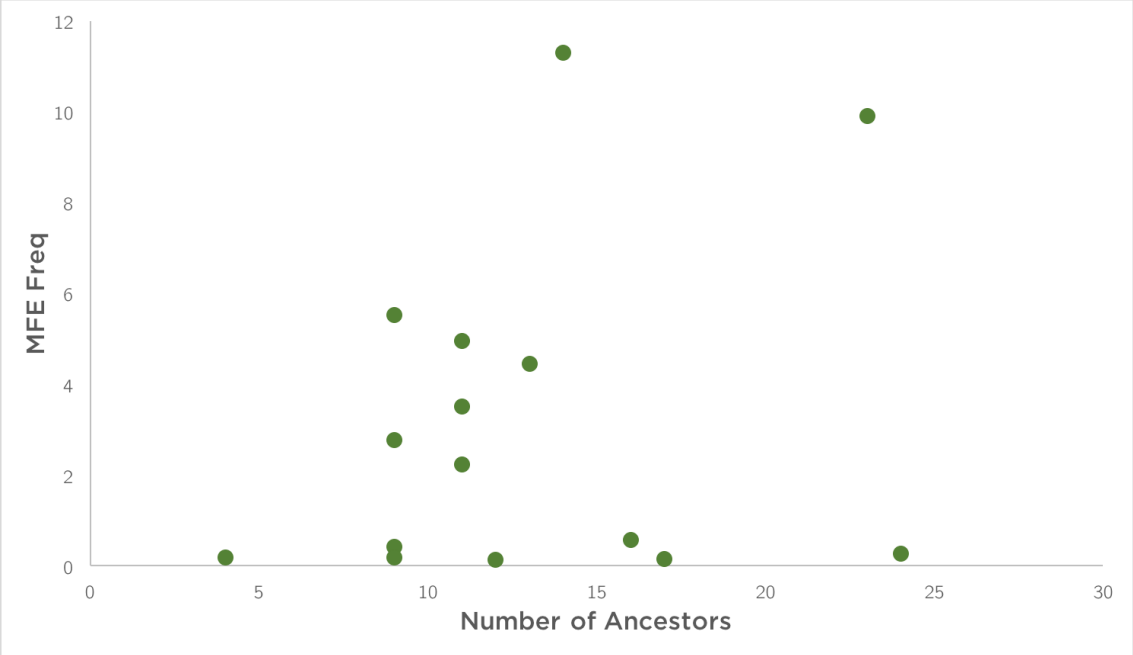

Figure 2: *Number of ancestors vs the net MFE frequency in an RNA family*

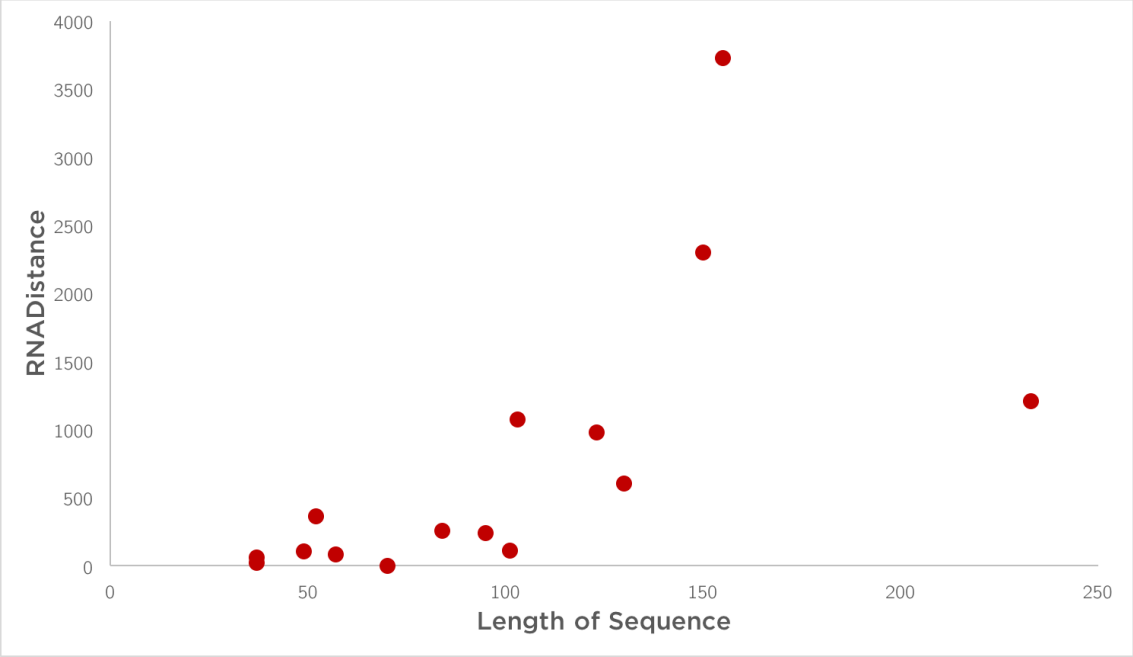

Figure 3: *Length of sequence vs the net RNADistance in an RNA family*

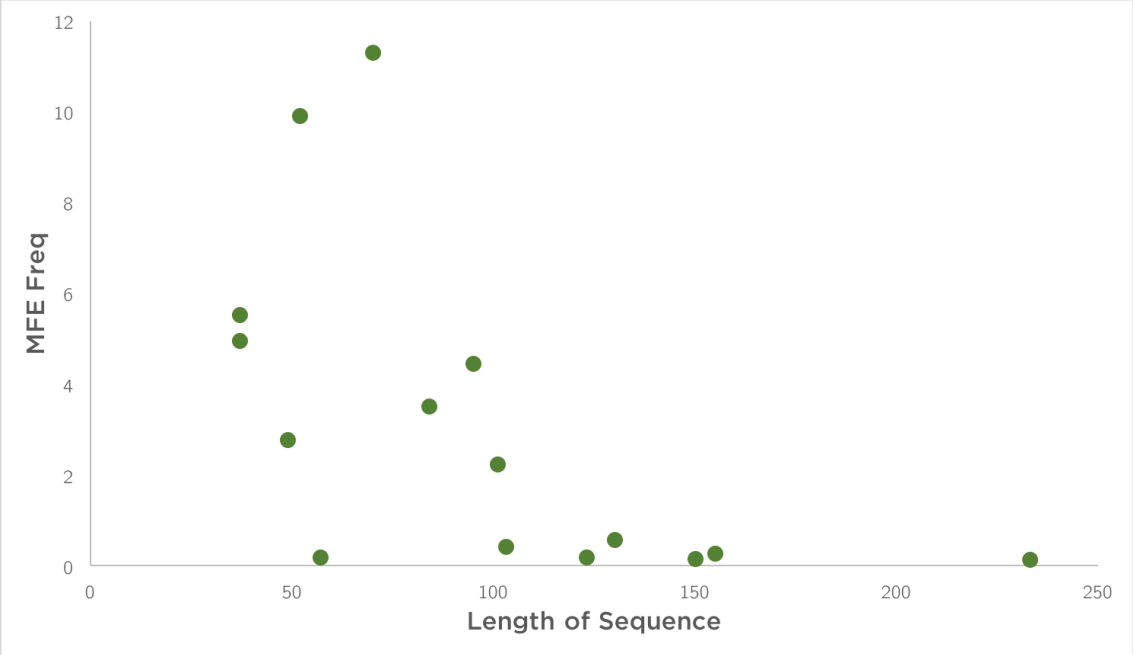

Figure 4: *Length of sequence vs the net MFE frequency in an RNA family*

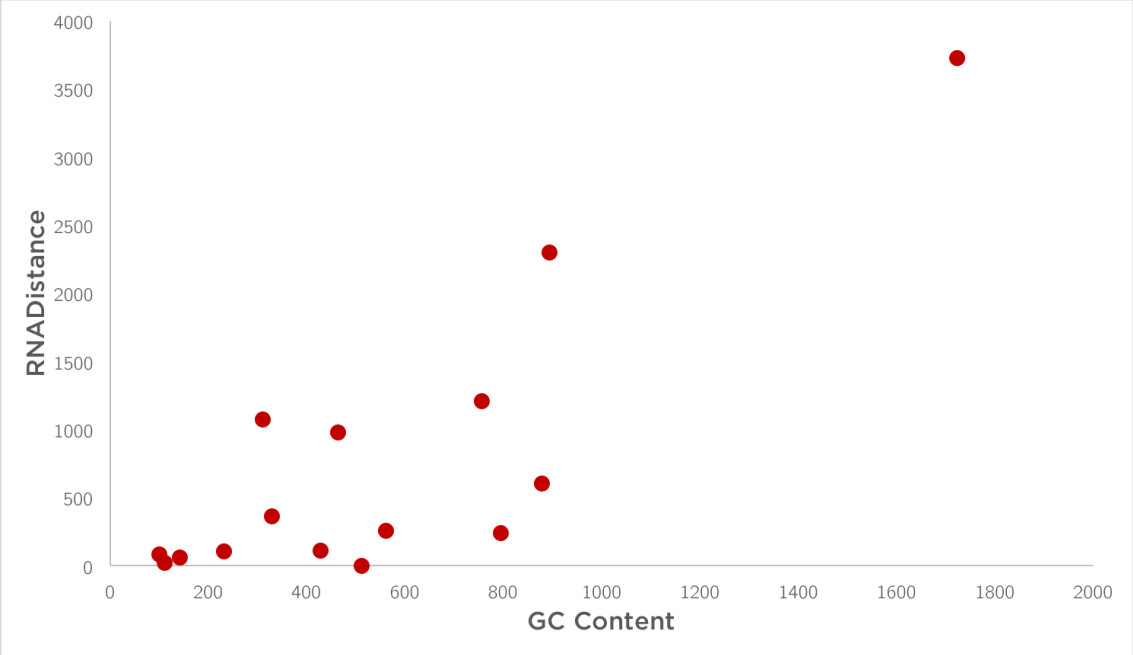

Figure 5: *GC content vs the net RNADistance in an RNA family*

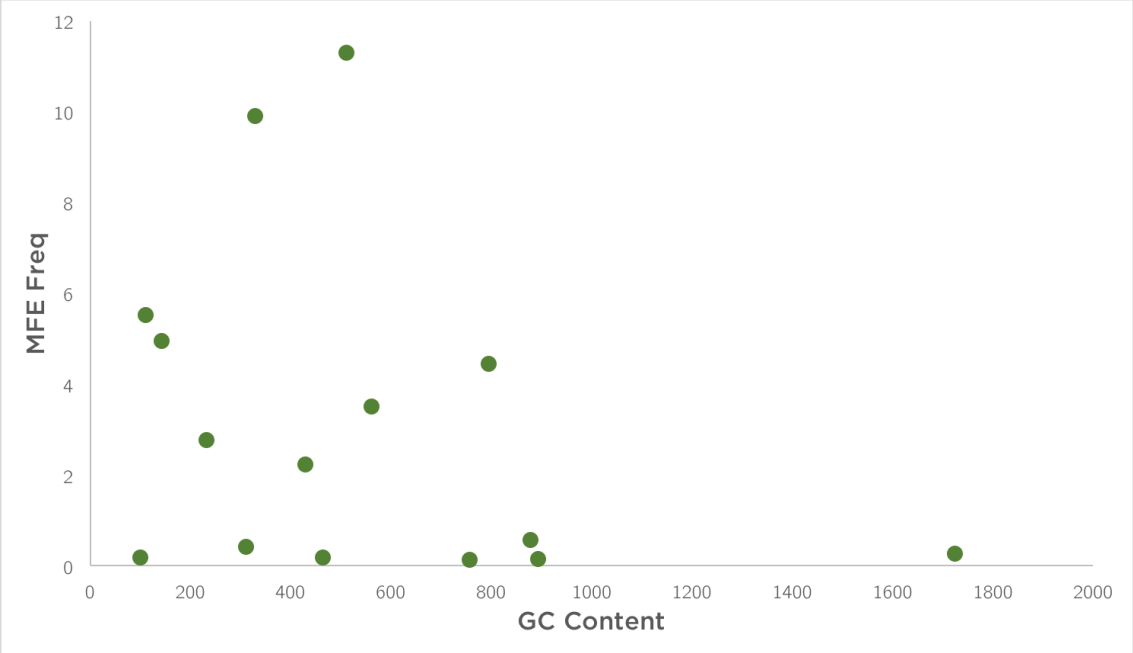

Figure 6: *GC content vs the net MFE frequency in an RNA family*

Discussion

We see that RNA Distance and the MFE frequency largely increases as the number of ancestors increase. We can say that they consensus structure dependencies are more conserved in s smaller phylogeny and the possibility of base pair mutations increases as the ancestral size increases.

For length of sequence, we find that longer sequences had a higher RNA Distance from the consensus structure. Also, shorter sequences tend to have higher MFE frequency.

GC content also behaves similarly. Higher GC content takes the secondary sequence away from the consensus structure and lower GC content generally has higher MFE frequency.

References

[1] Sankoff, David. Minimal mutation trees of sequences. SIAM Journal of Applied Mathematics 28:1 (1975) 35-42.

[2] Sankoff, David, and Pascale Rousseau. "Locating the vertices of a Steiner tree in an arbitrary metric space." Mathematical Programming 9.1 (1975): 240-246.

[3] Rfam: Home Page - http://rfam.xfam.org/

[4] Stockholm format - http://sonnhammer.sbc.su.se/Stockholm.html