Project Report

Comp 561 – Computational Biology Methods

Haji Mohammad Saleem

Table of Contents

Introduction

It is well known that there are millions of different species of organisms living on the earth today. Various studies have identified genetic similarities between different species, linking them together in an evolutionary relationship, discovered through molecular sequencing data and morphological data matrices. These related organisms could then become a part of a tree that tracks their evolutionary development and history by evaluating various common ancestors, representing the phylogeny of the organism.

The evolutionary history of any organism spans through ages, before leading to its present form. An evolved specie generally retains multiple ancestral features, gradually modified to aid in adaptability and survival. Thus a similarity in genetic data is a strong indicator of shared lineage and thus leads to common ancestors.

One of the modes of construction of 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. Fitch himself presented an algorithm known as Fitch's Algorithm to determine the cost of a given phylogeny.

We are interested in another algorithm by David Sankoff, which is a generalized version of the algorithm presented by Fitch. The algorithm itself is direct and can be studied in much detail here [1][2]. We would elaborate on the algorithm in the coming sections.

Our subjects for implementation of this algorithm are RNA families from the Rfam database hosted by the Wellcome Trust Sanger Institution collaboration with Janelia Farm. More information on the database is present here [3].

We access various RNA families annotated in the Stockholm format, which contains their sequence alignment and a consensus structure among other things. A detailed description of the Stockholm format is available here [4].

RNA molecules tend to fold onto themselves to create complex structures simplest of them being the secondary structure. It is the set of base pairs found in a single stranded RNA. Allowed base pairs are C-G, A-U (Watson-Crick) and G-U (Wobble). Many interesting RNA's conserve their secondary structure more than they conserve their sequence [5]. We thus predict the secondary structure of each ancestor and analyze its conservation in relation to the consensus structure.

Objectives

We had to complete several objectives as part of the project. They are as follows:

1. Stockholm File Parser to extract the sequence alignment and the consensus secondary structure.
2. Implementation of the Sankoff algorithm.
3. Expansion of the algorithm to account for gaps.
4. Application of the algorithm on an Rfam family.
5. Calculation of the secondary structure of each ancestor sequence and their distance.
6. Extension to conserve base pairing dependencies. (Note: obj 6 and 7 in the project description are merged into obj 6 in this document.)
7. Comparison between different Rfam families.

Tools

The tools used during the implementation of the project are:

1. Python – Python was used for the total implementation of the project.
2. Biopython – Library used to visualize the phylogenetic tree
3. Rfam – Database used to download Stockholm files of RNA families
4. Vienna RNA package – RNAfold and RNAdistance used to predict the secondary structure and distance from another structure.
5. Matploltib / MS-Excel – Used to visualize the comparison between different families.

Methods

- Objective 1: The objective required simple file handling and string manipulation. The result is written to standard output, which includes the sequence and consensus secondary structure. The lines that start without a # contain the sequences, each starting with its own individual name (used as a key to store in the dictionary) while the line with #=GC SS_cons saves the structure. Longer sequences may require splitting over multiple lines and were concatenated based on the key. The secondary structure required replacing "<" with "(" and ">" with ")" to match with the parenthesis notation (done in the later objectives).

- Objective 2: Implementation of the Sankoff algorithm. The algorithm read the tree provided manually in a file in the parenthesis notation and the sequences were provided manually as well. A manual cost matrix for different transitions was made as follows:

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

The tree sequence was converted into a dictionary of lists to create the hierarchical nature of the phylogenetic tree, eg. {1:[A,B]} represents a tree (A,B) with 1 as their ancestor. All the leaf nodes were provided with alphabetical codes for easy reference where as the ancestral nodes were provided numerical codes in the ascending order such that the oldest ancestor has the highest numerical code equal to one less than the number of sequences.

The weights for leaf nodes were initialized such that the nucleotide present has a weight 0 while the rest have a weight "inf". A list of lists was created to store the weights for each possible nucleotide (4 in this case) and for each index in the sequence (length of the sequence). Next we calculate the following for each index of the ancestral sequence with respect to the left and the right child:

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

where $S_a(i)$ is the smallest cost for node a for the state i. The states are chosen so as to minimize the cost to move to state j and k for left and right child.

This minimizing state forms a part of the ancestral sequence. We calculate the ancestors bottom up and finally at each index choose the nucleotide with the minimum cost(done in the later objectives). The result is printed on standard output.

- Objective 3: To expand our algorithm for gaps, we simply update our cost matrix for 5 symbols with a penalty of 2 for gap. The rest of the steps remain the same.

	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

- Objective 4: We choose RF00754 *microRNA mir-279*, to implement our algorithm. The family has 12 sequences of length 101 each.

  The tree sequence is manually provided in file and is as follows:

  ((((((A,B),C),(D,E)),F),(G,(H,I))),((J,K),L))

  The program directly reads the Stockholm file and runs our algorithm on it returning the ancestral sequences among other things in a text file. We share the output file in out results section.

- Objective 5: We use the functions RNAfold and RNAdistance from the Vienna RNA package to compute the secondary structure of each ancestor (11 in total) and its distance from the consensus secondary structure. We wrote a python wrapper for the functions to be called through the "os" library, adapted from a publically available version [6]. The output is stored in a text file shared in the results section.

- Objective 6: It is just an extension to the last objective. The aim is to conserve the base pair dependencies of the consensus secondary structure. The idea is to increase the possibility of a base pair at the same position as the consensus structure rather than the preserving the sequence itself. In order to do so, we keep two things in mind:
  a. The sequence index where either the base pair opens in the SS_Cons or closes should not have a gap and thus the next best nucleotide should be selected in case the gap has the least score.

b. The nucleotide at the closing index of a base pair should compliment the one at the opening. To tackle this we let the opening pair be chose by the algorithm (apart from the gap) and the closing pair be just nucleotide that forms a pair (C-G, A-U, G-U) and is the minimum cost in case of two possible.

The output is stored in a text file shared in the results section.

- Objective 7: Comparison of multiple families. We plan to compare almost 25 RNA families, calculating the net ancestral distance and comparing it with the number of seeds in the family, length of the sequence and net G-C content of the ancestors (I had emailed to clarify if we were to check the GC content of the ancestors). Since it would not be possible to manually provide a tree structure for each family, we would assume a gradually increasing tree for each family, eg., ((((A,B),C),D),E). We then produce the scatter plots for each comparison.

Results

Objective 1:

Input file: `RF00754_seed.stockholm.txt` (Stockholm file)
Standard Output

```
Consensus Secondary Structure

....<<<<<<<<<<<<<<<<<<<<<<<.<<<<<<....<<<<<.................>>>>>>>>>>>.>>>>>>>>>>>>>>>>>>>...>>..

Sequence Alignment

CAUCCCAACGGUUGUCGAUGGGUGUGA.AUCUAGUGUUUCACAUGAUUUUCGAUA.GCC..UGUGACUAGAU.CCACACUCAUUAACAAAAGUU...GCCG        A.aegypti.1
UCAUACUACUGUUUUUAGUGAGUGAGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCACACUCAUUAAUAACGGUA...GUUC        D.pseudoobscura.1
UACUACUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGAUUUCUG....UAUUUGUGACUAGAU.CCACACUCAUUAAUAACGGUA...GUUC        D.ananassae.1
CCGUAUUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGUUUAUUUG...UAUUUGUGACUAGAU.CCACACUCAUUAAAAACGGUA...GUUC        D.grimshawi.1
UCAUACUACUGUUUUUAGUGGGUGGGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCACACUCAUUAAUAACGGUA...GUUC        D.melanogaster.1
UUUCCUGAAUUUGCCAAAUGAGUGAAG.GUCUAGUG...CACAGAAAAUGAA.......AUUGUGACUAGAU.CCACACUCAUUAAGUACGUUC...AGGU        A.florea.1
UUGUUCGAUGGCCUUGGAUGGGUUUGA.AUUCAG...UCCACGUU..UUUUAUUU.UUAUUCGUGACUAGAU.CCACACUCAUCCAAGGAAAUC...GAGC        A.mellifera.1
GCUUCCCACUAUUGUCGAUGGGUGUGA.AUCUAGUGGUUCACAUGAGCUUUGCC..A.AACUGUGACUAGAU.CCACACUCAUUAACAAGAGUGCUCGGA        C.quinquefasciatus.1
AAUUUGAUCCGUUCUUGAUGGGUUCGG.GUCUAGUGG..CACGGUUUUUUCA.....ACUUCGUGACUAGAU.CCACACUCAUUAAGGAAGUUU...CACA        T.castaneum.2
UGGAGCUCUCGGUGUAAGCCAGUGUUCAGUCUAUUGUUUCACAUUGGUUUCG.......AUUGUGACUAGAUCGAACACUCGCUUGCAACCUGG...GUUU        T.castaneum.1
GCGUACUACUGUUUUUAGUGAGUGAGG.GUCCAGUGUUUCACAUUCGUUUUUUCAAGUAUUUGUGACUAGAU.CCACACUCAUUAACAACGGUA...GUUC        D.virilis.1
CCAGCCGAUUGUACUGAGUGAGUGAUG.GUCUGGUG.....CACGGUUUAUC.....GAUCUGUGACUAGAU.CCACACUCAUUAAGUACGUUC...GGCU        N.vitripennis.1
```

Objective 2,3 are intermediary to Objective 4:

Input file: `RF00754_seed.stockholm.txt` (Stockholm file)
`tree_rfam.txt` (tree sequence)
Output file: `RF00754out.txt` (file attached)

```
          Sankoff Algorithm on RNA Family RF00754
-------------------------------------------------------------------------------------------------------------------------
          Tree Sequence

          ((((((A,B),C),(D,E)),F),(G,(H,I))),((J,K),L))
-------------------------------------------------------------------------------------------------------------------------
          Tree Structure
                                                    _________ A
                                          _________|
                                _________|         |_________ B
                               |         |
                               |         |_________ C
                     _________|         |
                    |         |         |_________ D
                    |         |_________|
           _________|         |         |_________ E
          |         |         |
          |         |         |_________ F
 _________|         |
|         |         |_________ G
|         |_________|
|         |         |_________ H
|_________|         |_________|
|                   |_________ I
|
|         |_________ J
|         |
|_________|         |_________ K
          |
          |_________ L
```

```
          p          [lc, rc]

          1          ['A', 'B']
          2          [1, 'C']
          3          ['D', 'E']
          4          [2, 3]
          5          [4, 'F']
          6          ['H', 'I']
          7          ['G', 6]
          8          [5, 7]
          9          ['J', 'K']
          10         [9, 'L']
          11         [8, 10]
```

```
          Leaf Nodes

          A          T.castaneum.1
          C          D.virilis.1
          B          T.castaneum.2
          E          D.ananassae.1
          D          D.grimshawi.1
          G          D.pseudoobscura.1
          F          D.melanogaster.1
          I          C.quinquefasciatus.1
          H          A.aegypti.1
          K          A.mellifera.1
          J          A.florea.1
          L          N.vitripennis.1
```

```
          Leaf Sequence

          UGGAGCUCUCGGUGUAAGCCAGUGUUCAGUCUAUUGUUUCACAUUGGUUUCG.......AUUGUGACUAGAUCGAACACUCGCUUGCAACCUGG...GUUU        A
          GCGUACUACUGUUUUUAGUGAGUGAGG.GUCCAGUGUUUCACAUUCGUUUUUUCAAGUAUUUGUGACUAGAU.CCACACUCAUUAACAACGGUA...GUUC        C
          AAUUUGAUCCGUUCUUGAUGGGUUCGG.GUCUAGUGG..CACGGUUUUUUCA.....ACUUCGUGACUAGAU.CCACACUCAUUAAGGAAGUUU...CACA        B
          UACUACUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGAUUUCUG....UAUUUGUGACUAGAU.CCACACUCAUUAAUAACGGUA...GUUC        E
          CCGUAUUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGUUUAUUUG...UAUUUGUGACUAGAU.CCACACUCAUUAAAAACGGUA...GUUC        D
          UCAUACUACUGUUUUUAGUGAGUGAGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCACACUCAUUAAUAACGGUA...GUUC        G
          UCAUACUACUGUUUUUAGUGGGUGGGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCACACUCAUUAAUAACGGUA...GUUC        F
          GCUUCCCACUAUUGUCGAUGGGUGUGA.AUCUAGUGGUUCACAUGAGCUUUGCC..A.AACUGUGACUAGAU.CCACACUCAUUAACAAGAGUGCUCGGAA        I
          CAUCCCAACGGUUGUCGAUGGGUGUGA.AUCUAGUGUUUCACAUGAUUUUCGAUA.GCC..UGUGACUAGAU.CCACACUCAUUAACAAAAGUU...GCCG        H
          UUGUUCGAUGGCCUUGGAUGGGUUUGA.AUUCAG...UCCACGUU..UUUUAUUU.UUAUUCGUGACUAGAU.CCACACUCAUCCAAGGAAAUC...GAGC        K
          UUUCCUGAAUUUGCCAAAUGAGUGAAG.GUCUAGUG...CACAGAAAAUGAA.......AUUGUGACUAGAU.CCACACUCAUUAAGUACGUUC...AGGU        J
          CCAGCCGAUUGUACUGAGUGAGUGAUG.GUCUGGUG.....CACGGUUUAUC.....GAUCUGUGACUAGAU.CCACACUCAUUAAGUACGUUC...GGCU        L
```

```
          Cost Matrix

['A', 'C', 'G', 'U', '.']
[[0, 2, 1, 2, 2],
 [2, 0, 2, 1, 2],
 [1, 2, 0, 2, 2],
 [2, 1, 2, 0, 2],
 [2, 2, 2, 2, 0]]
```

```
          Calculated Ancestor Sequence

          AAGAGCACCCGGUCUAAACCAGUGCGCAGUCUAGUGGUUCACAGUGGUUUCA.....ACAUCGUGACUAGAUCCAACACUCACUAACAAACUGG...CACA        1
          AAGUACUACCGUUCUUAGUGAGUGAGG.GUCCAGUGUUUCACAUUCGUUUCAUCAAGAAUUUGUGACUAGAU.CCACACUCAUUAACAACGGUA...GUUC        2
          CACUACUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGAUUACUGG...UAUUUGUGACUAGAU.CCACACUCAUUAAAAACGGUA...GUUC        3
          UAGUACUACUGUUUUUAGUGAGUGAGG.GUCCAGUGUUUCACAUUGAUUUCUU....UAUUUGUGACUAGAU.CCACACUCAUUAACAACGGUA...GUUC        4
          UCAUACUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCACACUCAUUAAUAACGGUA...GUUC        5
          CAUCCCAACGAUUGUCGAUGGGUGUGA.AUCUAGUGGUUCACAUGAGCUUCGACA.ACAACUGUGACUAGAU.CCACACUCAUUAACAAAAGUGCUCGCAA        6
          CCAUACCACUGUUGUCAAUGAGUGAGA.AUCCAGUGUUUCACAUGAAUUUUCCUA.GCAACUGUGACUAGAU.CCACACUCAUUAACAAAAGUA...GCCA        7
          UCAUACUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCACACUCAUUAAUAACGGUA...GUUC        8
          UUGCCCGAAGGCCCCAAAUGAGUGAAA.AUCCAGUG.UCCACAGAAAAUGAAUUU.UUAAUCGUGACUAGAU.CCACACUCAUCAAAGAAAAUC...AAGC        9
          CCACCCGAUUGUACUGAAUGAGUGAAG.GUCUAGUG...CACACAAAUUAUA.....GAUCUGUGACUAGAU.CCACACUCAUUAAGUACGUUC...GGCU        10
          UCAUACGACUGUACUGAGUGAGUGAGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCACACUCAUUAAGAACGGUA...GGCC        11
```

Objective 5:

Input file: `RF00754_seed.stockholm.txt` (Stockholm file)
`tree_rfam.txt` (tree sequence)

Output file: `RF00754out3.txt` (file attached)

```
          RNAfold on Ancestral sequence of RNA Family RF00754
-------------------------------------------------------------------------------------------------------------------------
          Calculated Ancestor Sequence

          AAGAGCACCCGGUCUAAACCAGUGCGCAGUCUAGUGGUUCACAGUGGUUUCA.....ACAUCGUGACUAGAUCCAACACUCACUAACAAACUGG...CACA        1
          AAGUACUACCGUUCUUAGUGAGUGAGG.GUCCAGUGUUUCACAUUCGUUUCAUCAAGAAUUUGUGACUAGAU.CCACACUCAUUAACAACGGUA...GUUC        2
          CACUACUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGAUUACUGG...UAUUUGUGACUAGAU.CCACACUCAUUAAAAACGGUA...GUUC        3
          UAGUACUACUGUUUUUAGUGAGUGAGG.GUCCAGUGUUUCACAUUGAUUUCUU....UAUUUGUGACUAGAU.CCACACUCAUUAACAACGGUA...GUUC        4
          UCAUACUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCACACUCAUUAAUAACGGUA...GUUC        5
          CAUCCCAACGAUUGUCGAUGGGUGUGA.AUCUAGUGGUUCACAUGAGCUUCGACA.ACAACUGUGACUAGAU.CCACACUCAUUAACAAAAGUGCUCGCAA        6
          CCAUACCACUGUUGUCAAUGAGUGAGA.AUCCAGUGUUUCACAUGAAUUUUCCUA.GCAACUGUGACUAGAU.CCACACUCAUUAACAAAAGUA...GCCA        7
          UCAUACUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCACACUCAUUAAUAACGGUA...GUUC        8
          UUGCCCGAAGGCCCCAAAUGAGUGAAA.AUCCAGUG.UCCACAGAAAAUGAAUUU.UUAAUCGUGACUAGAU.CCACACUCAUCAAAGAAAAUC...AAGC        9
          CCACCCGAUUGUACUGAAUGAGUGAAG.GUCUAGUG...CACACAAAUUAUA.....GAUCUGUGACUAGAU.CCACACUCAUUAAGUACGUUC...GGCU       10
          UCAUACGACUGUACUGAGUGAGUGAGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCACACUCAUUAAGAACGGUA...GGCC       11
-------------------------------------------------------------------------------------------------------------------------
          Ancestor RNA Fold

          AAGAGCACCCGGUCUAAACCAGUGCGCAGUCUAGUGGUUCACAGUGGUUUCA.....ACAUCGUGACUAGAUCCAACACUCACUAACAAACUGG...CACA        1
          ....((((..(((....))).))))).((((.((((((((((((.(((.............))).)))).....------.))).))))).....))))........       84

          AAGUACUACCGUUCUUAGUGAGUGAGG.GUCCAGUGUUUCACAUUCGUUUCAUCAAGAAUUUGUGACUAGAU.CCACACUCAUUAACAACGGUA...GUUC        2
          ......(((((((.((((((((((.((.(((.(((....((.((((..........)))))..))..))).))).)).)).)))))))))).)))))).......       32

          CACUACUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGAUUACUGG...UAUUUGUGACUAGAU.CCACACUCAUUAAAAACGGUA...GUUC        3
          ......((((((((((((((((((.((.(((.((....(((((.................)))))))).))).)).)))))))))))))))))).......       16

          UAGUACUACUGUUUUUAGUGAGUGAGG.GUCCAGUGUUUCACAUUGAUUUCUU....UAUUUGUGACUAGAU.CCACACUCAUUAACAACGGUA...GUUC        4
          ......(((((((.((((((((((.((.(((.((....(((((.................)))))))).))).)).)))))))))).)))))).......       20

          UCAUACUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCACACUCAUUAAUAACGGUA...GUUC        5
          ......(((((((.((((((((((.((.(((.((....(((((.................)))))))).))).)).)))))))))).)))))).......       20

          CAUCCCAACGAUUGUCGAUGGGUGUGA.AUCUAGUGGUUCACAUGAGCUUCGACA.ACAACUGUGACUAGAU.CCACACUCAUUAACAAAAGUGCUCGCAA        6
          .......((..((((.((((((((((..((((((((((....)))))).................)))))))..))))))))))..)))).........       52

          CCAUACCACUGUUGUCAAUGAGUGAGA.AUCCAGUGUUUCACAUGAAUUUUCCUA.GCAACUGUGACUAGAU.CCACACUCAUUAACAAAAGUA...GCCA        7
          .......(((.((((.(((((((((.(..(((.((....(((((.................)))))))).))).)..).)))))))).)))).)))........       32

          UCAUACUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCACACUCAUUAAUAACGGUA...GUUC        8
          ......(((((((.((((((((((.((.(((.((....(((((.................)))))))).))).)).)))))))))).)))))).......       20

          UUGCCCGAAGGCCCCAAAUGAGUGAAA.AUCCAGUG.UCCACAGAAAAUGAAUUU.UUAAUCGUGACUAGAU.CCACACUCAUCAAAGAAAAUC...AAGC        9
          ..(((....))).....(((((((....(((.(((....(((.((.............)))))))).))).....------.)))))))..............       70

          CCACCCGAUUGUACUGAAUGAGUGAAG.GUCUAGUG...CACACAAAUUAUA.....GAUCUGUGACUAGAU.CCACACUCAUUAAGUACGUUC...GGCU       10
          ...((.((.((((((.(((((((((..(.(((((((.....((((.(((........))).)))))))))).)..)))))))).)))))).))...)).       38

          UCAUACGACUGUACUGAGUGAGUGAGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCACACUCAUUAAGAACGGUA...GGCC       11
          .......(((((.((.(((((((((.((.(((.((....(((((.................)))))))).))).)).)))))))).)).)))))........       28

-------------------------------------------------------------------------------------------------------------------------
```

The secondary structure is provided below the ancestor while its distance from the consensus structure is written below the index of the ancestor.

Objective 6:

Input file: `RF00754_seed.stockholm.txt` (Stockholm file)
`tree_rfam.txt` (tree sequence)

Output file: `RF00754out2.txt` (file attached)

```
          RNAfold on Ancestral sequence of RNA Family RF00754
-------------------------------------------------------------------------------------------------------------------------
          Calculated Ancestor Sequence

          AAGAGCACCCGGUCUAAACCAGUGCGCAGUCUAGUGGUUCACAGUGGUUUCA.....ACAUUGUGACUAGAUCGCACACUAAUUUAGACCGAGU...GUCA        1
          AAGUACUACCGUUCUUAGUGAGUGAGG.GUCCAGUGUUUCACAUUCGUUUCAUCAAGAAUUUGUGACUAGAU.CCUCACUCAUUAAGAACGGUA...GUUC        2
          CACUACUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGAUUACUGG...UAUUUGUGACUAGAU.CCUCACUCAUUAAAAACAGUA...GUUC        3
          UAGUACUACUGUUUUUAGUGAGUGAGG.GUCCAGUGUUUCACAUUGAUUUCUU....UAUUUGUGACUAGAU.CCUCACUCAUUAAAAACAGUA...GUUC        4
          UCAUACUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCUCACUCAUUAAAAACAGUA...GUUC        5
          CAUCCCAACGAUUGUCGAUGGGUGUGA.AUCUAGUGGUUCACAUGAGCUUCGACA.ACAACUGUGACUAGAU.UCACACUCAUUAACAAUCGUUCUCGGAA        6
          CCAUACCACUGUUGUCAAUGAGUGAGA.AUCCAGUGUUUCACAUGAAUUUUCCUA.GCAACUGUGACUAGAU.UCUCACUCAUUAACAACAGUA...GUCA        7
          UCAUACUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCUCACUCAUUAAAAACAGUA...GUUC        8
          UUGCCCGAAGGCCCCAAAUGAGUGAAA.AUCCAGUG.UCCACAGAAAAUGAAUUU.UUAAUUGUGACUAGAU.UUUCACUCAUUUAAGACCUUC...AAGC        9
          CCACCCGAUUGUACUGAAUGAGUGAAG.GUCUAGUG..CCACACAAAUUAUA.....GAUCUGUGACUAGAU.CUUCACUCAUUCAGUACAAUC...GGCU       10
          UCAUACGACUGUACUGAGUGAGUGAGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCUCACUCAUUCAGUACAGUC...GUCC       11
-------------------------------------------------------------------------------------------------------------------------
          Ancestor RNA Fold

          AAGAGCACCCGGUCUAAACCAGUGCGCAGUCUAGUGGUUCACAGUGGUUUCA.....ACAUUGUGACUAGAUCGCACACUAAUUUAGACCGAGU...GUCA        1
          ......((.(((((((((..((((.((.((((((....(((((((.............)))))))))))))).)).)))))..)))))))))..)).......       36

          AAGUACUACCGUUCUUAGUGAGUGAGG.GUCCAGUGUUUCACAUUCGUUUCAUCAAGAAUUUGUGACUAGAU.CCUCACUCAUUAAGAACGGUA...GUUC        2
          ......((((((((((((((((((((.(((.(((....((.((((...........)))))..))..))).))).))))))))))))))))))).......       24

          CACUACUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGAUUACUGG...UAUUUGUGACUAGAU.CCUCACUCAUUAAAAACAGUA...GUUC        3
          ......((((((((((((((((((((.(((.((....(((((.................)))))))..))).))))))))))))))))))))).......       12

          UAGUACUACUGUUUUUAGUGAGUGAGG.GUCCAGUGUUUCACAUUGAUUUCUU....UAUUUGUGACUAGAU.CCUCACUCAUUAAAAACAGUA...GUUC        4
          ......((((((((((((((((((((.(((.((....(((((.................)))))))..))).))))))))))))))))))))).......       12

          UCAUACUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCUCACUCAUUAAAAACAGUA...GUUC        5
          ......((((((((((((((((((((.(((.((....(((((.................)))))))..))).))))))))))))))))))))).......       12

          CAUCCCAACGAUUGUCGAUGGGUGUGA.AUCUAGUGGUUCACAUGAGCUUCGACA.ACAACUGUGACUAGAU.UCACACUCAUUAACAAUCGUUCUCGGAA        6
          ..(((.(((((((((.((((((((((.((((((((((....)))))..................)))))))).)))))))))).)))))))))...))).       34

          CCAUACCACUGUUGUCAAUGAGUGAGA.AUCCAGUGUUUCACAUGAAUUUUCCUA.GCAACUGUGACUAGAU.UCUCACUCAUUAACAACAGUA...GUCA        7
          .......((((((((.((((((((((.(((.((....(((((.................)))))))..))).))))))))))).)))))))).........       20

          UCAUACUACUGUUUUUAGUGGGUGAGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCUCACUCAUUAAAAACAGUA...GUUC        8
          ......((((((((((((((((((((.(((.((....(((((.................)))))))..))).))))))))))))))))))))).......       12

          UUGCCCGAAGGCCCCAAAUGAGUGAAA.AUCCAGUG.UCCACAGAAAAUGAAUUU.UUAAUUGUGACUAGAU.UUUCACUCAUUUAAGACCUUC...AAGC        9
          ......(((((.(..((((((((((((.(((.(((....((((.................)))))))..))).))))))))))))..).)))))).......       24

          CCACCCGAUUGUACUGAAUGAGUGAAG.GUCUAGUG..CCACACAAAUUAUA.....GAUCUGUGACUAGAU.CUUCACUCAUUCAGUACAAUC...GGCU       10
          ...((.(((((((((((((((((((.(((((((......((((.(((........))).)))))))))).)))))))))))))))))))...))..       22

          UCAUACGACUGUACUGAGUGAGUGAGG.GUCCAGUGUUUCACAUUGAUUUUCUUA.GUAUUUGUGACUAGAU.CCUCACUCAUUCAGUACAGUC...GUCC       11
          ......((((((((((((((((((((.(((.((....(((((.................)))))))..))).))))))))))))))))))))).......       12

-------------------------------------------------------------------------------------------------------------------------
```

The secondary structure is provided below the ancestor while its distance from the consensus structure is written below the index of the ancestor.

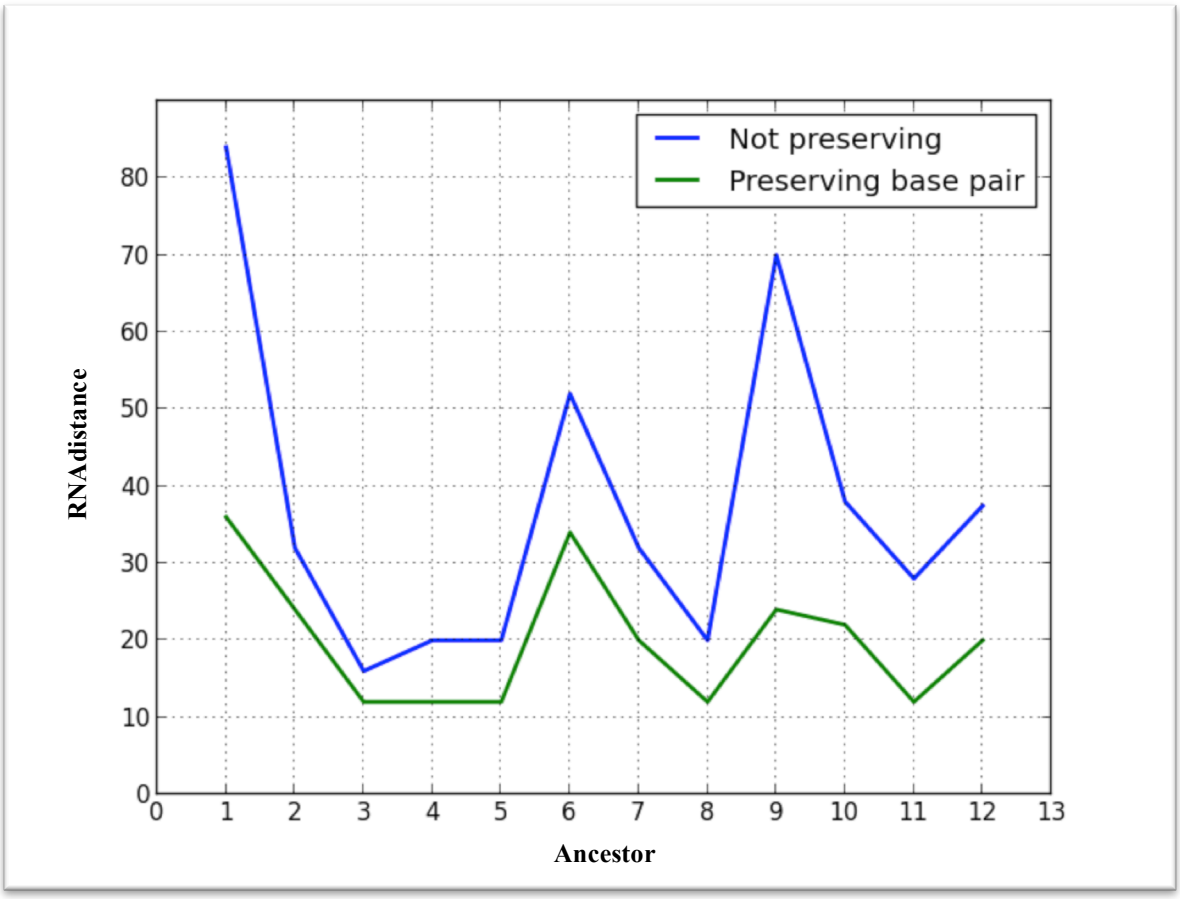

Fig 1: *RNAdistance for each ancestor in Objective 5 and 6.*
The twelfth point represents the average distance.

Objective 5-6 Discussion:

We can clearly see from figure 1 that the RNAdistance for every ancestor goes down when we conserve base-pair dependencies of the consensus secondary structure.

This leads to lowering of the average distance by almost 45%.

Thus, conserving the base pair dependencies keeps the ancestral sequences closer to the consensus secondary structure.

Objective 7:

Input file: `multiple Stockholm files`

Output file: `compare.txt` (converted to a table presented here)

Name	Num of Seqs	Len of Seq	Distance	GC content
RF00196	6	35	0	54
RF01065	19	116	98	684
RF01313	10	57	156	235
RF00754	12	101	158	446
RF01066	20	97	170	1130
RF00131	49	64	244	1290
RF00240	16	77	292	411
RF00374	23	104	298	1290
RF00951	31	52	460	444
RF00128	17	176	622	1252
RF00761	44	109	624	1331
RF01803	52	54	666	1470
RF00694	28	111	690	1275
RF00246	32	92	720	1141
RF00667	50	80	876	1707
RF00252	18	190	880	1275
RF02131	22	139	1020	1765
RF01769	25	129	1194	1371
RF00668	47	79	1196	1309
RF00130	41	111	1340	1810
RF00654	36	98	1608	1445
RF00083	21	249	1810	1994
RF01497	26	118	2072	1676
RF00022	27	284	2526	2161
RF01502	51	754	6828	10683

Table 1: *Number of sequences, length of sequence, RNAdistance and GC content of various RNA families*

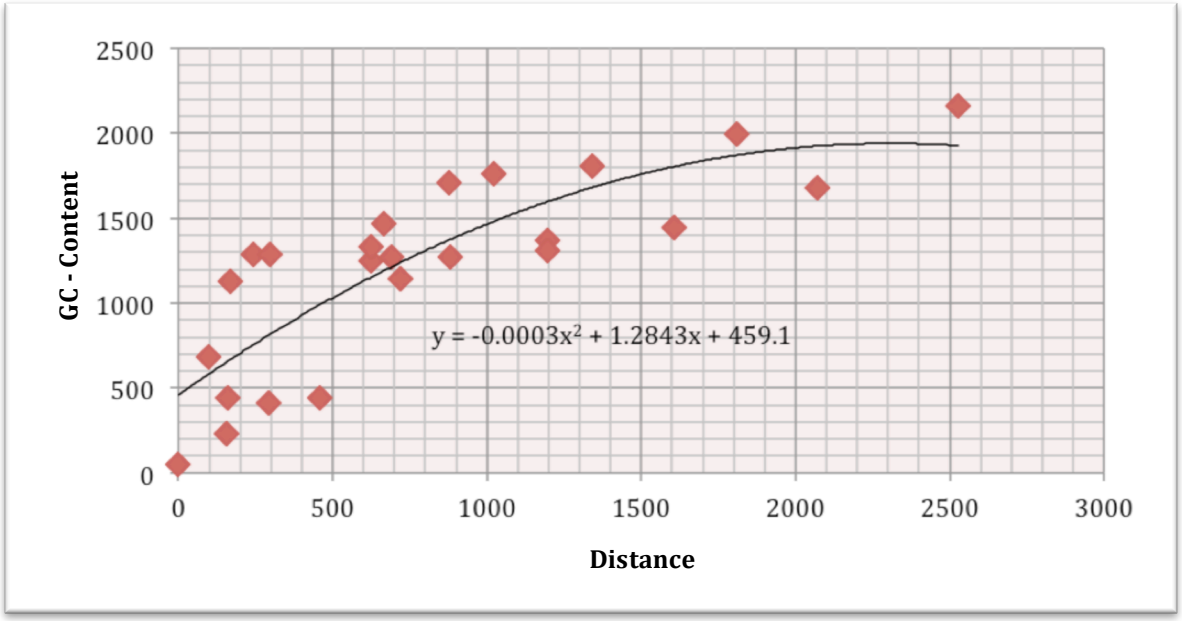

Fig 2: *net RNAdistance of all ancestors vs GC content of all the ancestors.*

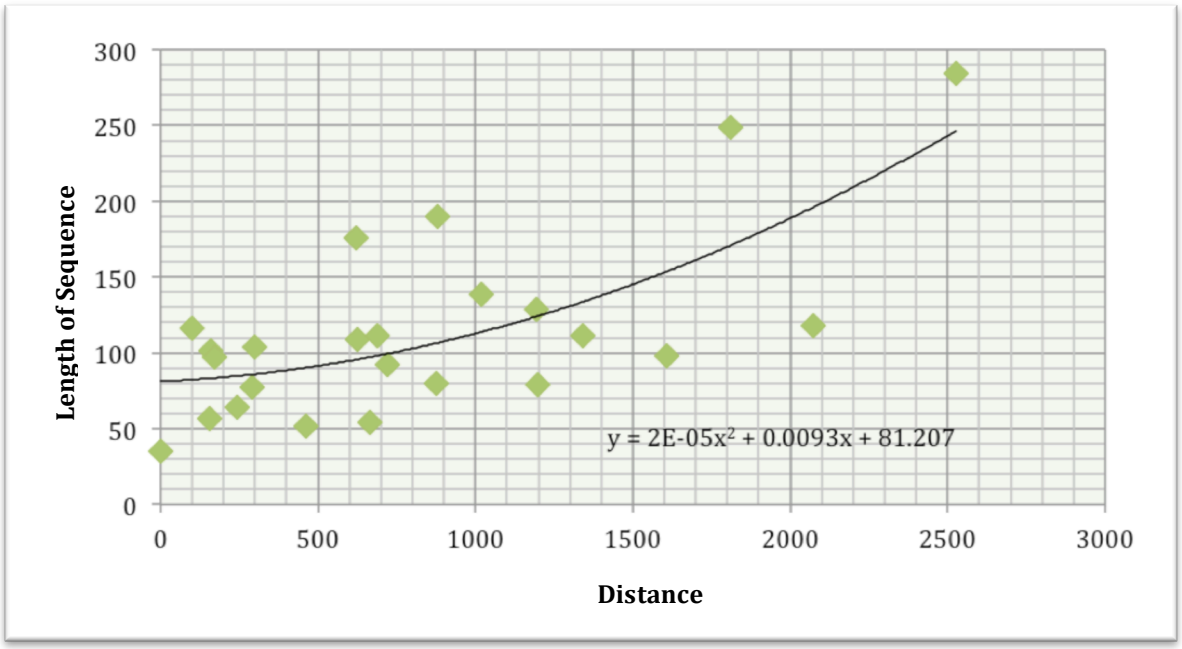

Fig 3: *net RNAdistance of all ancestors vs the length of sequence of the family.*

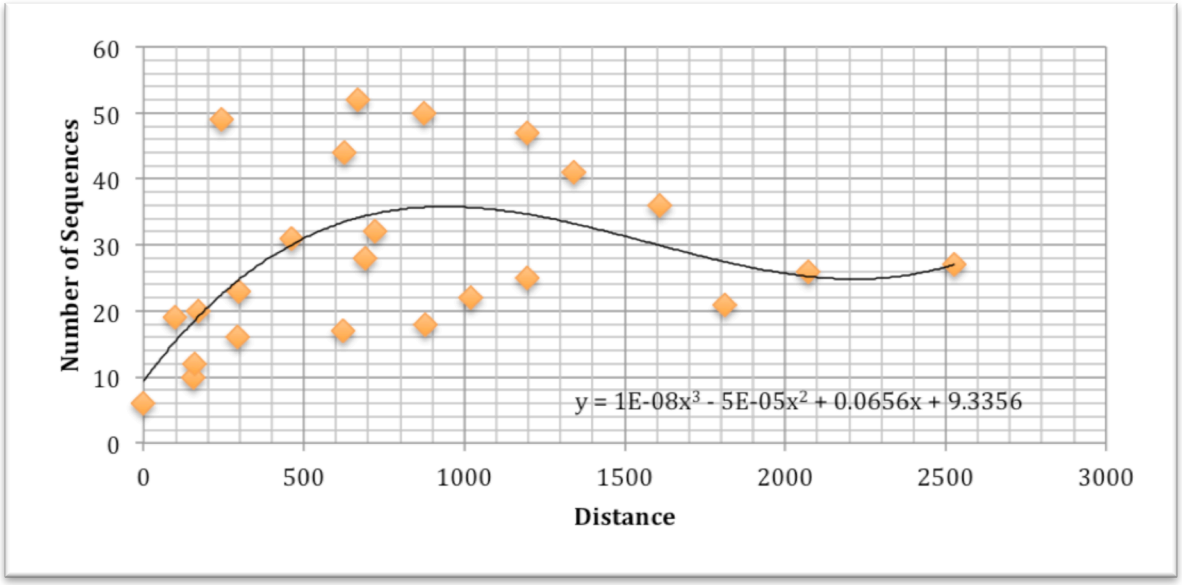

Fig 4: *net RNAdistance of all ancestors vs the number of sequences in a family*

Objective 7 Discussion:

We wish to understand the relation between various features of a family and its ability to conserve the secondary structure in the ancestral sequences.

We use the RNAdistance as an inverse proxy measure for the degree of conservation of consensus structure.

Our three graphs present the following results:

1. GC content and distance have a relationship characterized by a polynomial of degree 2. Distance increases at a much higher rate as the GC content increases. This is most likely due to the fact a lot more strong base pairs are formed, than the ones present in the secondary structure, due to the abundance of the G and C nucleotides.
2. A longer sequence relates to a higher distance through a polynomial of degree 2, though the rate is not as high as GC-content. This is most likely due to the fact that more indices increases the chance of more base pairs but not as quickly as the presence high number of GC nucleotides and thus increasing the distance because of higher number of pairs.
3. Number of sequences does not seem to have a totally increasing or decreasing relationship and depends on other factors. For example a lot of sequences with the smaller length are able to conserve the structure much better than less sequences with longer sequence.

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.sanger.ac.uk*

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

[5] Durbin, Richard, ed. *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge university press, 1998.

[6] Adapted from Vienna Wrappers for Python by Yann Ponty – *http://www.lix.polytechnique.fr/~ponty/enseignement/ViennaWrappers.py*

Appendix

The python scripts corresponding to different objectives are as follows:

- Objective 1 – Stockholm_parser.py
- Objective 2 – sankoff_algo.py
- Objective 3 – sankoff_algo_withgaps.py
- Objective 4 – sankoff_on_rfam.py
- Objective 5 – rnafold_ancestor.py
- Objective 6 – rnafold_ancestor_preserve.py
- Objective 7 – rfam_compare.py

Github link to the project: https://github.com/hajisaleem/comp561