RNA ancestors
Comp 598 project

Paul Pereira

December 19, 2016

1 Introduction

The RFAM database gives us access to a multitude of non coding RNA families. Non-coding RNA's are considered to belong to the same family if there secondary structure and therefore function is the same. The sequence of non-coding RNA's belonging to the same family can however be very different. In this project, we will study the utility of the Sankoff algorithm [] in predicting RNA ancestor sequences of RNA families. We will propose a modification of the algorithm that takes into account a consensus structure when predicting the ancestor sequences and we will look at different metric that can be used to determine the validity of our predictions. Finally, we will look at possible reasons for our modified algorithm to fail at providing an improvement over the original method.

2 Method

2.1 Sankoff Algorithm

The sankoff algorithm is a modified version of the Fitch algorithm this uses weighted mutation cost matrix in order to calculate the most likely ancestor sequence between two RNA sequences. The sankoff algorithm takes as input a set of RNA sequences belonging to the same RNA family. The sankoff algorithm also takes as input the structure of the parsimony tree for that family, representing the evolutionary distance between the sequences in our set. Each sequence in our set represents a leaf in the tree. An internal node in the tree always has two children.
The objectif of the algorithm is to calculate the most likely ancestral sequence for every internal node in our parsimony tree. In the case of the Sankoff algorithm, a cost is associated with every mutation between the ancestral sequence and the sequences of the two children nodes. The RNA sequences are made up of 5 characters Σ = (A,C,G,U,-) representing the four possible nucleotides or a gap in the sequence. The cost matrix used is:

Table 1: 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

Keeping the same nucleotide from the children to the ancestor sequence is free. A substitution from a purine to another purine or a pyrimidine from another pyrimidine has a cost of 1. Anything else adds a cost of 2. The method to calculate the sequence is dynamic. Let $C_{s,i,n}$ be the cost of having nucleotide n at position i in sequence s. If s is a leaf in the tree (i.e a sequence in the set provided as input) then $C_{s,i,n} = 0$ if i is the nucleotide at position n in s and ∞ otherwise. In the case of an internal node, let s_1 and s_2 be the two children sequences of s, then:

$$C_{s,i,n} = min_{j \in \Sigma}(C_{s_1,j,n} + cost(i,j)) + min_{k \in \Sigma}(C_{s_2,k,n} + cost(i,k)) \quad (1)$$

With the score for each nucleotide for each position in the sequence computer, the ancestral sequence is computed by choosing at every position the nucleotide with the lowest score :

$$s_n = argmin_{k \in \Sigma} C_{s,k,n} \quad (2)$$

2.2 RNAfold and RNAdistance

Once the sequence has been computed, we used the RNAfold software from the Vienna RNA package to determine the minimum free energy secondary structure of the sequence. The RFAM database contains family of RNA sequences that are non-coding. Their main function therefore will be to interact with other RNA molecules or protein. The function of the RNA sequence is determined by its structure. The MFE structure gives us the most likely (stable) confirmation that the RNA molecule can take.

We know furthermore that RNA sequences belonging to the same family will perform similar functions and as such have a similar structure. The stockholm format from the RFAM database gives us an a consensus secondary structure for every family. It would make sense for the ancestor sequences that we are trying to predict to have a structure similar to the consensus sequence (the ancestor of two sequences carrying out the same function must have had the same function as its two children). This provides us with a way to measure the efficacy of the Sankoff Algorithm. We can argue that a prediction for an ancestor sequence is good only if the MFE structure of that sequence is similar to the consensus sequence.

We use the RNAdistance software from the Vienna RNA package to compare the predicted MFE structure of an ancestor sequence to the consensus structure.

2.3 Using the consensus secondary structure to modify Sankoff

We suggest a way to improve the prediction of the Sankoff algorithm. Instead of simply using the consensus structure as a way to evaluate our prediction, we can try to modify our prediction so as to make it fit the consensus structure. A new sequence is obtained in the following way : for every base pair positions (n,m) in the consensus structure check if the nucleotides at position i and j in our predicted sequence can base pair (i.e are either A-U, C-G or G-U). If it is the case, do nothing. Otherwise, find the min (u*,v*) over all base pairing pair of nucleotides (u,v) of $C_{s,u,n} + C_{s,v,m}$ and change the nucleotides at position n and m in the sequence to u* and v*. This modification does not ensure that the predicted ancestor sequence will have an MFE structure similar to the consensus sequence. However, it does guarantee that every base pair in the consensus sequence can occur during the folding of the predicted sequence. The modified algorithm picks the pair (u*,v*) that minimizes the combined cost of having nucleotide u* at position n and nucleotide v* at position m.

3 Results

We tested both the original and modified version of Sankoff on 3 RNA families. In addition to calculating the MFE structure and the distance of that structure to the consensus sequence, we also used RNAfold to determine the stability of each MFE structure and looked at how this value evolved based on the length of the sequences in each family, the CG content and which version of the Sankoff algorithm was used.
The following table provides a summary of the 3 family studied (the name of the families, the length of each sequence in the alignment and the number of sequences studied)

Table 2: Summary

Family Name	length	N sequences
chlorobi	70	9
frnS	136	8
RprA	120	8

The following table provides a detailed description of the results obtained for the RprA family. The columns with an extra N represent the results obtained using the modified version of our algorithm. The id of each row can be used to identify the corresponding node in the structure of the parsimony tree. The remaining results can be found in the appendix.

Table 3: RprA results

id	distance	distance N	CG content	CG N	stability	stability N
0	61	53	0.425	0.425	0.0385133	0.0339571
1	43	65	0.316666667	0.308333333	0.0222112	0.086275
2	43	43	0.291666667	0.283333333	0.079729	0.0359312
3	53	49	0.266666667	0.258333333	0.135568	0.109959
4	43	43	0.291666667	0.283333333	0.079729	0.0359312
5	61	45	0.25	0.25	0.0693992	0.0295158
6	51	43	0.325	0.316666667	0.00465963	0.00934491
7	81	85	0.391666667	0.375	0.117877	0.0590128
8	69	51	0.416666667	0.416666667	0.00436339	0.00996885
9	41	49	0.3	0.275	0.0475319	0.024732
10	39	35	0.25	0.241666667	0.0947445	0.095204
11	73	41	0.291666667	0.275	0.169524	0.0941774
12	53	29	0.3	0.291666667	0.0259072	0.00751461
13	69	63	0.4	0.375	0.0263495	0.0585413
14	59	57	0.35	0.35	0.205494	0.0680651

4 Discussion

We will now discuss the efficiency of the algorithm and its modification.

The first observation that we can make is that the length of the sequences has a significant effect on the quality of the results. In the case of the chlorobi-1 family, the sequence are 70 nucleotides long(short). We observe that the distance obtained with RNAdistance ranges from 1 to a maximum of 13 which is small. Furthermore, the MFE structure is very stable, with the sequence at the root obtaining a score of 0.83. Results vary a bit more on other internal nodes based on who the children of the nodes are. For example, node 12 has a stability of 0.06 due to each children node 4 and 5 having a low stability score (0.27 and 0.5). In addition, in the case of a small sequence, we observe that the modification we suggest for Sankoff greatly improve the results. The distance between the MFE and the consensus is almost always 1. Furthermore, we see an increase in the stability of the MFE structures predicted. The stability of the MFE structure of node 12 for example goes from 0.06 to 0.6. We also observe some MFE structures that have a lower stability (for example node 16). This suggests that on sequences of small sizes, modifying the predicted sequence so as to maintain the base pairing properties defined in the consensus sequence can help promote ancestral sequences with more stable MFE structures.

On the hand, in the case where we have longer sequences, such as with the frnS and RprA family, using our modified algorithm does not seem to improve the stability of the structures we obtain.

For example, the stability scores in the RprA family ranges from 0.004 to 0.2 (the smaller values compared to the chlorobi-1 familoy can be explained by the fact that with a a longer sequence, there are more opportunities to form structures of similar free energy). In the case of the structures obtained when using the modified version however, most of the stability scores are smaller and range from 0.009 to 0.1. The distance from the consensus sequence is not improved significantly either except for node 12. One interesting observation is that a shortening the distance with the consensus structure does not necessarily improve stability. For example, the distance for node 12 goes from 53 to 29 and yet, the stability score decreases when using modified Sankoff.

We know that one possible contributing factor that might skew the results is the CG content of each sequence. We know that C-G base pairs are more stable than A-U base pairs are therefore we might be observing higher stability scores for sequences with higher CG content because of their high CG content. This indeed seems to be the case when we compare the chlorobi results to the frnS results. The CG content for the chlorobi family ranges from 0.47 to 0.57 whereas for the frnS family, the CG content ranges from 0.26 to 0.35. However, it is important to note that inside the families, a higher CG content doe snot automatically correlate with a higher stability score. For example in the chlorobi parsimony tree node 6 has a CG content of 0.5 but a stability score of only 0.42. Therefore, it is most likely that the difference observed between the RNA families is due to the length of the sequence.

5 Conclusion

In conclusion, in this project we have looked at applying the sankoff algorithm in order to predict ancestor sequences of RNA families. We have proposed a modification of this algorithm in order to take into account the consensus structure provided by the RFAM database. We have shown that this modification is useful when the length of the sequences is short and therefore the number of possible secondary structure is small. We have also shown that when sequences get longer, the modification does not improve our results. In the future, it would be interesting to consider another way of integrating the consensus structure into the Sankoff algorithm, perhaps by modifying the C matrix when we perform a change to the sequence so that the changes are reflected in the calculation of other ancestor sequences.

References

[1] David Sankoff *Minimal mutation trees of sequences* 1975.

6 Appendix

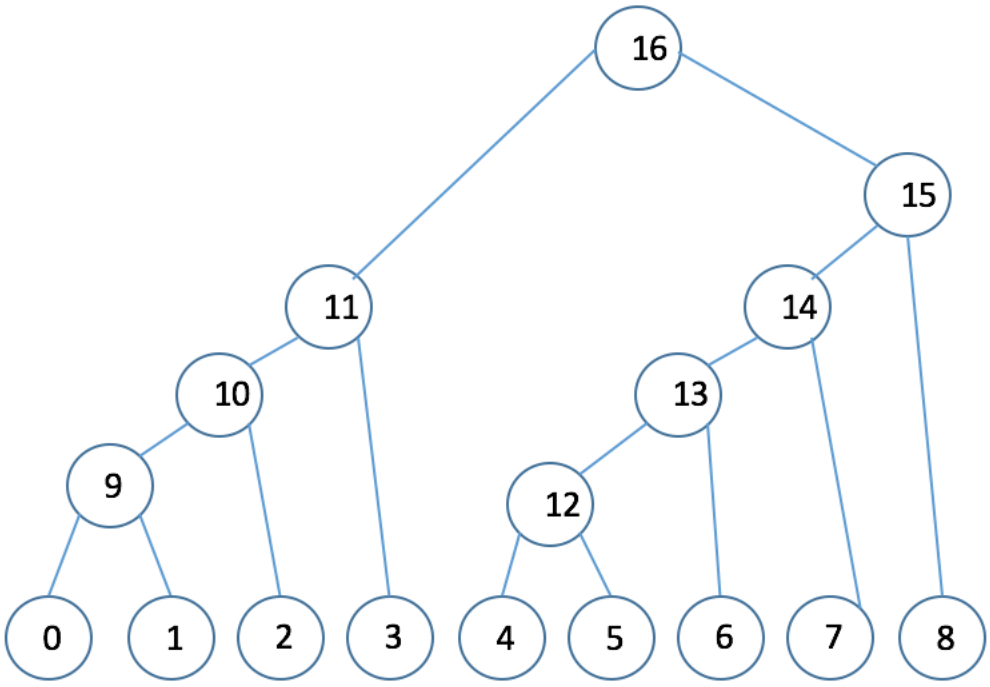
Table 4: chlorobi results

id	distance	distance N	CG	CG N	stability	stability N
0	5	1	0.542857143	0.542857143	0.715332	0.932608
1	5	1	0.542857143	0.542857143	0.715332	0.932608
2	5	1	0.514285714	0.514285714	0.783273	0.951075
3	9	1	0.5	0.5	0.684559	0.656948
4	9	1	0.442857143	0.428571429	0.27429	0.585828
5	1	1	0.571428571	0.571428571	0.503544	0.503544
6	5	1	0.5	0.485714286	0.423687	0.387613
7	1	1	0.542857143	0.542857143	0.697671	0.697671
8	9	5	0.585714286	0.571428571	0.487335	0.494256
9	5	1	0.542857143	0.542857143	0.715332	0.932608
10	5	1	0.528571429	0.528571429	0.775872	0.942093
11	13	1	0.485714286	0.471428571	0.770803	0.953668
12	15	1	0.485714286	0.457142857	0.0608173	0.600397
13	13	1	0.5	0.471428571	0.35135	0.412778
14	5	1	0.485714286	0.471428571	0.705099	0.789215
15	9	1	0.557142857	0.542857143	0.873693	0.758747
16	5	1	0.528571429	0.528571429	0.836073	0.790046

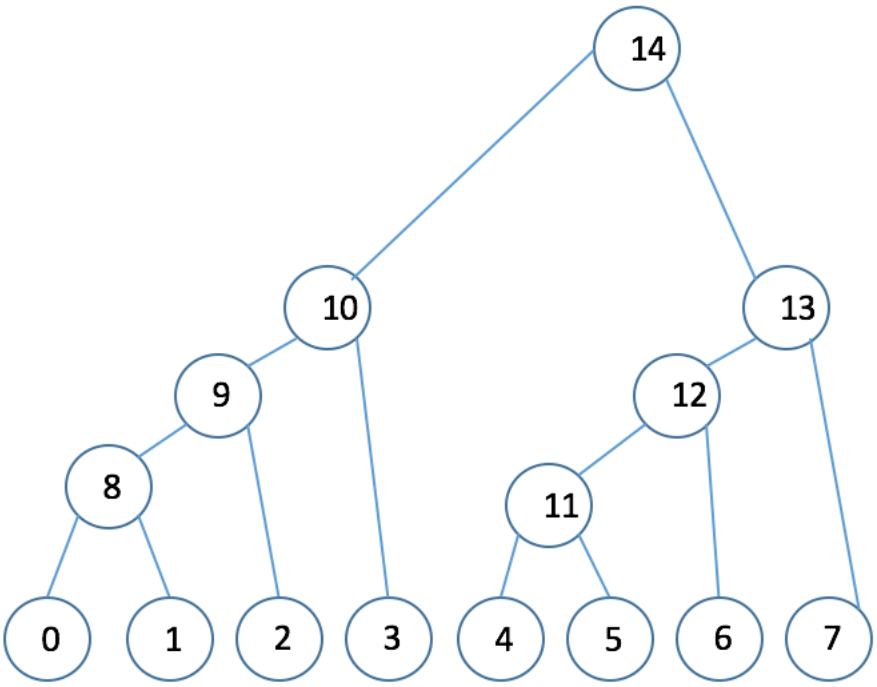
Table 5: frnS Results

id	distance	distance N	CG	CG N	stability	stability N
0	41	35	0.367647059	0.323529412	0.0133523	0.0350837
1	61	31	0.345588235	0.323529412	0.00640704	0.0456405
2	47	39	0.360294118	0.338235294	0.0102936	0.104796
3	99	11	0.345588235	0.316176471	0.00403391	0.0171684
4	77	47	0.338235294	0.308823529	0.00867286	0.00364013
5	69	59	0.308823529	0.286764706	0.0177857	0.0116131
6	81	51	0.345588235	0.308823529	0.00820782	0.00337815
7	81	61	0.301470588	0.264705882	0.0275064	0.0683962
8	61	27	0.352941176	0.352941176	0.00784716	0.0249831
9	37	25	0.360294118	0.352941176	0.0331038	0.0837428
10	81	31	0.338235294	0.338235294	0.0119401	0.0185174
11	83	25	0.330882353	0.286764706	0.00884965	0.00757132
12	81	51	0.330882353	0.301470588	0.00859332	0.0054698
13	69	61	0.323529412	0.286764706	0.0209884	0.02146
14	79	15	0.330882353	0.316176471	0.0170338	0.0066932

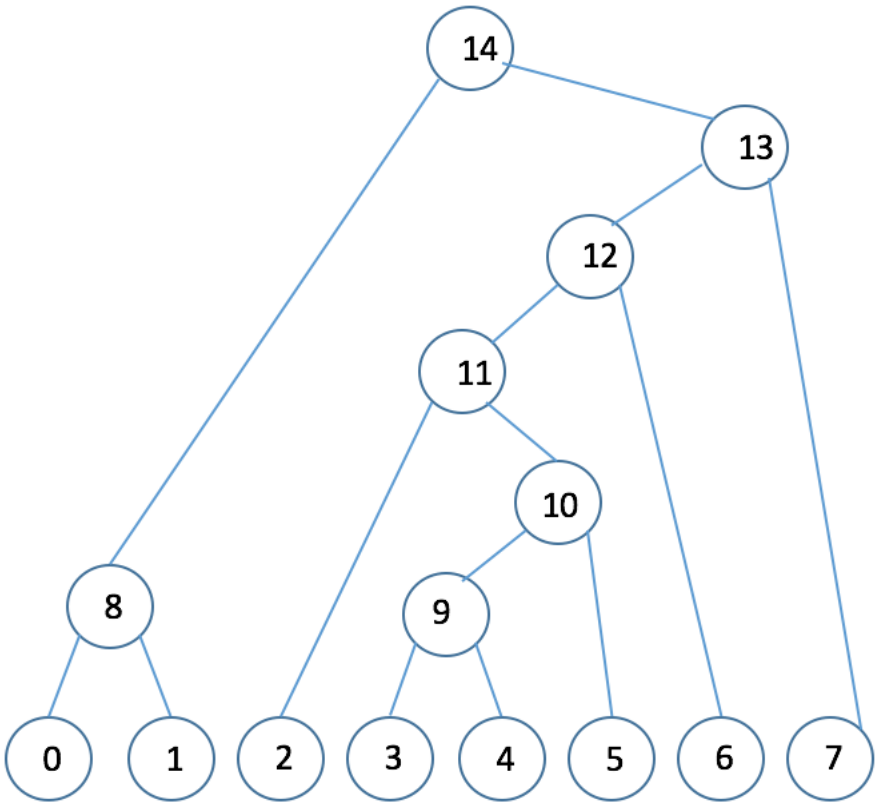
The parsimony tree for the chlorobi-1 family.



The parsimony tree for the frnS family.



The parsimony tree for the RprA family.


14
13
12
11
10
8
9
0
1
2
3
4
5
6
7