Master Degree Project

UNIVERSITY OF SKÖVDE

Investigation of phylogenetic relationships using microRNA sequences and secondary structures

Author:
Rohit Dnyansagar
a07rohdn@student.his.se

Supervisor:
Angelica Lindlöf
angelica.lindlof@his.se

School of Life Sciences
Skövde University
BOX 408
SE-541 28 Skövde
Sweden

Investigation of phylogenetic relationships using microRNA sequences and secondary structures

Rohit Dnyansagar

Master's dissertation

University of Skövde

2010

Investigation of phylogenetic relationships using microRNA sequences and secondary structures

Rohit Dnyansagar

Submitted by Rohit Dnyansagar to the University of Skövde as dissertation towards the degree of Master by examination and dissertation in the School of Life Sciences.

2010

I certify that all material in this thesis which is not my own work has been identified and that no material is included for which a degree has previously been conferred on me.

_____

Rohit Dnyansagar

Abstract

MicroRNAs are important biomolecules for regulating biological processes. Moreover, the secondary structure of microRNA is important for its activity and has been used previously as a mean for finding unknown microRNAs. A phylogenetic study of the microRNA secondary structure reveals more information than its primary sequence, because the primary sequence can undergo mutations that give rise to different phylogenetic relationships, whereas the secondary structure is more robust against mutations and therefore sometimes more informative.

Here we constructed a phylogenetic tree entirely based on microRNA secondary structures using tools *PHYLIP* (Felsenstein, 1995) and *RNAforester* (Matthias Höchsmann, 2003, Hochsmann et al., 2004), and compared the overall topology and clusters with the phylogenetic tree constructed using microRNA sequence. The purpose behind this comparison was to investigate the sequence and structure similarity in phylogenetic context and also to investigate if functionally similar microRNA genes are closer in their structure-derived phylogenetic tree.

Our phylogenetic comparison shows that the sequence similarity has hardly any effect on the structure similarity in the phylogenetic tree. MicroRNAs that have similar function are closer in the phylogenetic tree based on secondary structure than its respective sequence phylogeny. Hence, this approach can be very useful in predicting the functions of the new microRNAs whose function is yet to be known, since the function of the miRNAs heavily relies on its secondary structure.

Introduction

MicroRNAs belong to the non-coding RNAs having a regulatory function (Alvarez-Garcia and Miska, 2005). They are ~22 nucleotides long. Recently microRNAs have received much attention because of the recent genome-wide analyses, which have revealed that microRNAs are widely spread among a variety of species and also, are evolutionary conserved (Altuvia et al., 2005). Considerably, microRNAs have now been established as one of the largest gene families. MicroRNAs target and repress the activity of nearly 60% of all genes in an organism (Friedman et al., 2009). MicroRNAs are continuously being discovered along with their target genes. Since their discovery, both cloning methods and computational approaches have been used to identify new microRNA genes as well as to process relevant data to further understand the function of the microRNAs.

In animal cells, the mature microRNA sequences are obtained from pre-miRNA (about 70-100 nucleotides), which fold into a stem loop structure that is essential for the maturation process. Initially, Drosha, the nuclear ribonuclease III, cuts the pre-miRNA and then pre-miRNAs are exported to the cytosol by the exportin-5 pathway. Thereafter cytoplasmic ribonuclease III dicer excises the mature miRNA gene (Wang et al., 2005). The maturation process of microRNAs in plant cells is similar but the lengths of pre-miRNAs are more variable and with more complex structures (Wang et al., 2005). Additionally, the mechanism of action of the microRNAs differs in plants and animals. In plants the microRNA gene perfectly binds their target thereby causing degradation of the mRNA, while microRNAs in animals prevent translation without mRNA degradation. The exact mechanism by which bound microRNAs down-regulate target mRNA translation is currently unknown.

Phylogenetics is a widely used approach by biologists for the investigation of the hypothesis about the evolutionary path followed by group of organisms (Congdon, 2006). A typical phylogenetic study involves the derivation of a phylogenetic tree, wherein the root of the tree can be considered as a common ancestor of the species included in study. There are many methods to derive phylogenies, e.g., parsimony, maximum likelihood and MCMC-based Bayesian inference etc. (Congdon, 2006). Comparison of species or gene sequences in phylogenetic context can provide very important understanding to biology, such as reveal evolutionary patterns of morphological and chemical characters as well as many complex pathways. Molecular phylogenetics is a branch of phylogenetics that makes use of the molecular structure to derive phylogeny and attempts to determine the rates and patterns of change in the molecule. The overall idea is to interpret process from patterns. Non-coding functional RNA molecules are a major data source in molecular phylogenetics. The characteristics of RNA evolution and its long

term conservation of secondary structure have recently been considered for phylogenetic study (Primer, 2004).

With the rapidly increasing number of structures of biological molecules, a major challenge for biologist is to infer the function of these biological molecules based on their structural similarity (Primer, 2004). Rather than sequence, the structure of a molecule can give more prominent evidence of its evolutionary as well as functional role in biology (Liang and Landweber, 2005). For example, two proteins can have very low sequence similarity, but if they are similar on the structural level this is an indication of functional similarity. In case of non-coding RNAs, their function is mediated by their secondary structure. Since the structural conservation is not in accordance with the sequence conservation, the primary sequence is more mutable or exchangeable without or with very small consequences on its secondary structure. Therefore phylogeny based on primary sequence can give variable results and interpretations, whereas the structure provides a more robust mean to derive the phylogenetic relationships among species. In this study we derive the phylogeny of microRNAs entirely based on their secondary structure to investigate if functionally related microRNA sequences present any structural similarity.

Secondary structure of a RNA molecule is caused due to the hydrogen bonds that are formed between the base pairs; there are two types of hydrogen bonds (Lu, 2010): a) Watson and crick bonds and b) wobble bonds. Secondary structure is determined by Watson and crick (G≡C, A=U) and wobble (G−U) base pairing between the bases. A thermodynamic consideration of the secondary structure is that the structure having minimum free energy can be considered as the actual secondary structure of the RNA molecule. During base pairing the Watson crick pairing is stronger than the wobble base pairing. Base pairing stabilizes the structure whereas loops and unpaired bases destabilize the structure (Lu, 2010).

Problem Description

A general consideration is that sequence similarities often implicate structural similarity. Although this statement is true in most of the cases we cannot be sure about the ancestry or its relationship with all molecular structures. MicroRNA genes face some evolutionary constraints (Tanzer and Stadler, 2006) in order to recognize its target - the mature microRNA sequence has to be unchanged since it targets multiple genes and therefore the condition of co-evolution along with target genes is less likely to occur.

There are previous approaches wherein the microRNA structure was used to improve the phylogeny or to identify new microRNAs from the genome (Wang et al., 2005, Berezikov et al., 2005).This investigation aims to check whether if there is any correlation between the phylogeny study based on the sequence data and study based on the structure data. Also in this study we investigated if the functionally related microRNAs have any correlation between them.

The accuracy of the molecular phylogeny depends on the quality of the alignments made. Molecular systematics mostly relies on the computers to do the alignments. But the RNA molecules are prone to indel mutations and therefore at time of the alignment gaps and gap penalties must be taken into consideration to get a good alignment (Kjer, 1995).Therefore the mutations do affect the alignment of sequences and phylogenetics thereof. On the other hand the secondary structure of the molecule is more robust to mutations and therefore can give more reliable alignments and phylogenetic information. In this study we expect to find differences in the phylogenetic relationships based on the microRNA secondary structure and sequence, respectively.

A similar structure often implies similar function, the reverse is seldom true. Recent research shows that the microRNAs show conservation in the sequence. However, in general when the functional RNA families are observed, commonly sequence conservation is not found but structure conservation exists on the other hand (Hochsmann et al., 2004). Therefore, the structural study about the conservation and phylogeny of microRNA becomes a topic of interest since it may better represent functional relatedness (e.g. targeting the same type of genes). Second aspect to investigate in this study is if the functionally similar microRNAs are more closely related in phylogenetic tree based on secondary structure.

Further, the study is focused on the investigation of the phylogenetic relationship of the microRNAs that are causing diseases or malfunction. So here our goal is also to investigate if microRNAs that are responsible for a particular disease or malfunction are more closely related to each other structurally than in their sequences.

Material and methods

Dataset and selection

The source for all microRNA sequences is the miRBase 14.0 (Griffiths-Jones, 2006) where all the published microRNA entries are collected, and available for browsing and downloading in Fasta, XML or HTML file format. For the current study the Fasta format is used and the following organisms are selected randomly: *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster* and *Caenorhabditis elegans*. While considering the biogenesis of microRNAs and different forms during the process, i.e. pri-miRNA, pre-miRNA and mature miRNA, only pre-miRNA has a secondary structure and therefore this form is used for the purpose of the study. In case of microRNAs, the mature sequences are very similar and conserved. Thus, if we construct a phylogeny based on these sequences, we would get a tree wherein all sequences are closely related. From those trees one may infer that these genes are responsible for the same function, however when studying the structure this may not be true.

The downloaded dataset contains all microRNA sequences for all organisms available in the database in one single file. Therefore the sequences for the selected organisms are extracted from that file using a Python script.

For the further analysis a second database, miR2Disease (Jiang et al., 2009) is also used, which contains manually curated microRNA genes that are responsible for causing a specific disease or malfunction in the human body. For the current study the microRNA genes that are responsible for causing lung and prostate carcinoma and Colorectal Neoplasm are considered. The database contains a list of microRNA genes which are causing a particular disorder and microRNA genes which have some unspecified role in the disorder. Therefore for this particular study the genes which are responsible for causing the disorder are selected.

Sequence analysis

For the sequence analysis two conventional tools are used, the first one being ClustalW v1.83 (Thompson et al., 1994), used for sequence alignment and the other being PHYLIP (Felsenstein, 1995), that is used for phylogenetic evaluation.

Initially, the sequences extracted from the multifasta file are aligned using ClustalW and then the alignment file obtained is fed to another application called *distmat* (Carver, 2010), which gives the distance matrix of the aligned sequences in PHYLIP acceptable format. The distance matrix obtained from the *distmat* (Carver, 2010) program is used as input to PHYLIP's *neighbor* program which uses the Neighbor-Joining algorithm to produce a tree file from the distance

matrix. The tree file obtained from the previous step is viewed and analyzed using the application *Dendroscope* (Huson et al., 2007).

Structure analysis

For the structural analysis the Vienna RNA package is used which is a compilation of programs for RNA Secondary Structure Prediction and Comparison. It is Linux-based and for that purpose the Linux environment DNALinux (Bassi, 2007) is used. Among many programs in the Vienna RNA package two programs are of particular interest for this study, the first one is *RNAfold* and other one is *RNAforester*.

RNAfold (Zuker and Stiegler, 1981, McCaskill, 1990) is a command line based tool for calculating the RNA secondary structure. *RNAfold* reads in the RNA sequence from the standard input and calculates the structure based on *minimum free energy* and prints it to the standard output (Figure 1B). It also produces post script files containing the secondary structure graph (Figure 1A) and a dot plot of base pairing matrix. There are many programs having different approaches to predict RNA secondary structure.

RNAforester (Matthias Höchsmann, 2003, Hochsmann et al., 2004) is a program that calculates the similarity between two or multiple RNA secondary structures (i.e. pairwise and multiple secondary structure comparison). The input required for this program is the Vienna (DotBracket) format (Figure 2) which usually is the output from the *RNAfold* program.

Experiments

Here the aim is to investigate whether the sequence similarity of miRNAs is expressed in structure similarity in phylogenetic contexts. To investigate this we need to compare the two phylogentic trees for miRNAs based on sequence and the secondary structure. The sequence tree of the microRNAs from an organism is obtained by aligning the sequences using crustalW (Thompson et al., 1994) and then making a phylogenetic tree using PHYLIP's (Felsenstein, 1995) neighbor program. PHYLIP's neighbor program is combined implementation of the two clustering methods first is Neighbor-Joining and other is UPGMA. By successive clustering of lineages, PHYLIP constructs phylogenetic tree. Branch length is set as the lineages join.

The structure tree of the microRNAs from an organism is obtained by first predicting the microRNA secondary structure by the RNAfold (Zuker and Stiegler, 1981) program included in the Vienna RNA package. Then the secondary structures are aligned by the RNAforester

(Hochsmann et al., 2004, Matthias Höchsmann, 2003) program. The phylogenetic tree is obtained by PHYLIP neighbor program

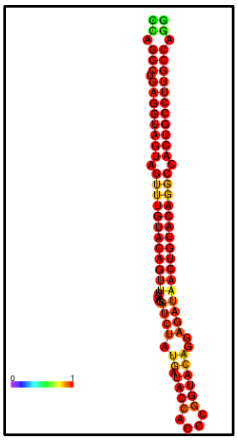
The topological similarity between the two trees is calculated using the 'Pairwise comparison of phylogenies'.

In order to investigate if the functionally related microRNAs are close in the secondary structure based phylogenetic trees than in sequence based phylogenetic trees the above procedure for the comparison of the phylogenetic trees is performed on the gene sequences which are responsible for causing disease or malfunction.

Results

In the earlier sections we discussed why phylogeny based on secondary structures will be interesting to study. In this particular study we present a new approach for the study of phylogeny based entirely on the secondary structure.

Many of the programs that predict RNA secondary structure consider formation of pseudoknots (Staple and Butcher, 2005). As this study is mainly focused on microRNAs which are small and have a simple hairpin structure, simple programs like RNAfold which predicts the secondary structure by minimum free energy are sufficient. Similarly, there are other programs that can be used for comparing the RNA secondary structure. But, since input format required for the RNAforester (Hochsmann et al., 2004) and the output obtained is in coordination with other programs in the Vienna RNA package, it is easier to use. RNAforester also gives a pairwise similarity score, which is helpful when creating a similarity profile.



```
Input string (upper or lower case); @ to quit
....,....1....,....2....,....3....,....4....,....5....,....6.
UACACUGUGGAUCCGGUGAGGUAGUAGGUUGUAUAGUUUGGAAUAUUACCACCGGUGAAC
UAUGCAAUUUUCUACCUUACCGGAGACAGAACUCUUCGA
length = 60
UACACUGUGGAUCCGGUGAGGUAGUAGGUUGUAUAGUUUGGAAUAUUACCACCGGUGAAC
............((((((..(((((((..(((.......)))..))))))))))))).....
 minimum free energy = -13.80 kcal/mol
```

(A) (B)

Figure 1. (A) Typical postscript image with base pair probabilities from the RNAfold program showing base pairing in the RNA molecule. This is a representation that is useful for the forest representation of the RNA molecule and thereby is a method of comparison. **(B)** Output from the RNAfold showing the secondary structure prediction in DotBracket format along with the minimum free energy value.

```
*** Scoring parameters ***

Scoring type: similarity
Scoring parameters:
pm:    10
pd:    -5
bm:    1
br:    0
bd:    -10


Input string (upper or lower case); & to end for multiple alignments, @ to quit
....,....1....,....2....,....3....,....4....,....5....,....6....,....7....,....8
global optimal score: 52
hsa-let-7a-1              UGGGAUGAGGUAGUAGGUUGUAUAGUUUUAGGGUCACACCCAC-CA-CUGGG-AG
hsa-let-7a-2              AGGU-UGAGGUAGUAGGUUGUAUAGUUU-AG----A-A---UUACAUCAAGGGAG
                            **  ********************** **    * *      ** *  ** **
hsa-let-7a-1              AUAACUAUACAAUCUACUGUCUUUCCUA
hsa-let-7a-2              AUAACUGUACAGCCUCCUAGCUUU-CCU
                          ****** ****  ** **  **** *

hsa-let-7a-1              (((((.(((..((((((((((((((((...(((.....))).(-(.-...))-..
hsa-let-7a-2              (((.-.(((.(((.((((((((((((.-..----.-.---(..(.....)..).
                          ***  ***** ** ************** *     * *      ** ****   *
hsa-let-7a-1              ..))))))))))))))))..)))))))))
hsa-let-7a-2              ..))))))))))))).)))).)))).-)))
                          *************** ** ****  ***
```

Figure 2. : RNAforester output which shows the alignment of the input secondary structure and score for the alignment. RNAforester also gives the pairwise similarity scores.

Initially the secondary structure of the selected genes is predicted using RNAfold algorithm and then these secondary structures are given as input to RNAforester to compare the secondary structures. RNAforester results in pairwise similarity score between each of the input genes as well as consensus sequence and the consensus secondary structure. All the default parameters of RNAforester are used for the comparison. From the output file the similarity score are separated and inserted into a new file. Using a perl script these similarity score are converted into a similarity profile. This profile is formatted manually in order to be of format suitable for the PHYLIP *neighbor* program. PHYLIP neighbor program uses the Neighbor-joining algorithm to convert the similarity scores to a phylogenetic tree. This phylogenetic tree is visualized and analyzed using the program dendroscope (Huson et al., 2007).

1) Comparison between the species tree

The following organisms are selected for the analysis: *Homo sapiens (721 genes), Drosophila melanogaster (157genes), Caenorhabditis elegans (174 genes),* and *Mus musculus (579 genes).* By comparison it can be seen that the sequence similarity does not necessarily implicate the structural similarity in phylogeny. As observed in the Figure 3 the comparison of the two topologies gives about 20.8 % similarity score calculated by the *Pairwise comparison of the phylogenies* (Nye et al., 2006). The thin lines indicate a better match between the two trees and the thicker lines indicate a worse match. As we can observe (Figure 3) there are plenty of the thicker lines indicating that there are not much of a similarity between the tree generated using the sequence data and the tree generated using structure data. Other topological comparison for the drosophila melanogaster is given in the Appendix A. The phylogenetic trees for the *Homo sapiens* and *Mus musculus* are too large to compare with this algorithm thus the topological similarity for them is not included in this study.

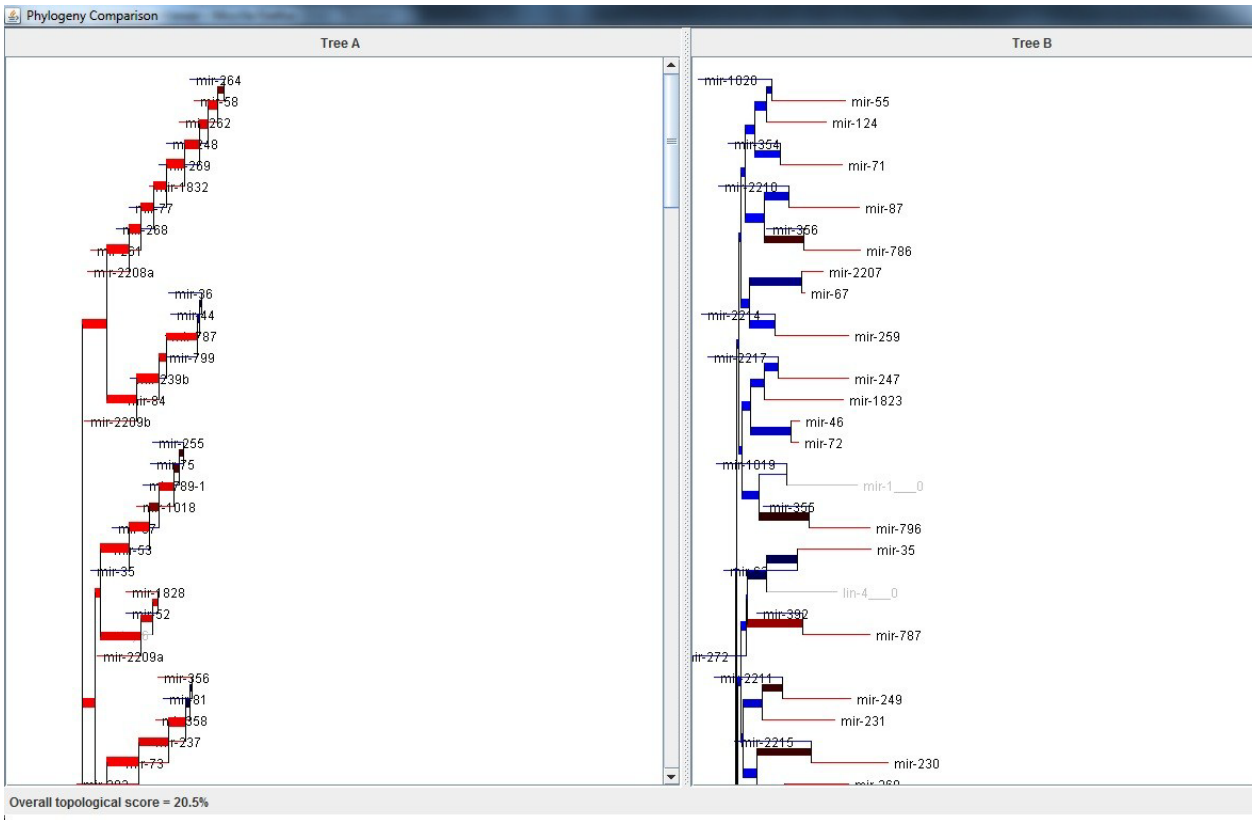


Figure 3. Comparison of the topology of the two phylogenetic trees generated for the organism *Caenorhabditis elegans*. On the left side there is tree generated using sequence data and on the right side there is tree generated using structure data.

2) Comparison between the cancer causing microRNA genes

A further study is done to check the phylogenetic relationship between the microRNA genes that are responsible for causing diseases or malfunctions. For the current study genes responsible for carcinoma of lung, carcinoma of prostate and colorectal neoplasm's are separated and phylogenetic relationship of single disease and also a combination study is done. The entire focus of the study is to compare the phylogenetic trees produced using microRNA sequences and microRNA secondary structure. There are many disease related data in the database miR2Disease. To perform a study covering all these disease related data was out of the scope of time duration. So study covering some disease related data selected randomly is presented here.

Study 1:

Figure 4 and figure 5 show the trees resulted from the microRNA sequence and microRNA structure respectively. Some genes involved in this study are responsible for causing colorectal carcinoma and some are responsible for causing prostate cancer. The C

suffix after the gene name indicates that the gene is responsible for causing colorectal carcinoma and P suffix after the gene indicates that the gene is responsible for causing the prostate carcinoma. When the tree resulted from sequence and tree resulted from secondary structure are compared for their topological similarity, it shows only 30.4% topological similarity to each other.

The purpose behind this study is not to study the ancestry of microRNA genes but to study the relationship between the microRNAs. So we can ignore the nodes and the roots of the phylogeny and can concentrate on the leaves of the phylogenetic tree. In Figure 4 and Figure 5 we can see phylogenetic trees for the study involving genes causing colorectal cancer and prostate cancer. In the Figure 5 we can observe that microRNAs that are responsible for causing a disease are more closely related to each other than in Figure 4. This is from the observation that there are more number of pairs in Figure 5 whose function is same. So genes responsible for the similar function are closely related in phylogenetic tree based on the secondary structure than phylogenetic tree based on sequence. Method of comparison of the phylogenetic trees is limited to visual inspection due to the fact that, even if the of phylogenetic tree construction algorithm is same in both the trees the scoring function in the distance matrix preparation is different.

A similar trend is observed in the remaining two studies (the results are given in Appendix A).

Study 2:
In this case the genes selected for the study are responsible for causing colorectal carcinoma and lung carcinoma. Suffix C indicates that the gene is responsible for causing colorectal carcinoma and L indicates lung carcinoma. The resulting trees i.e. the trees generated using the sequences and the secondary structure of those genes share 30.1 % topological similarity.

Study 3:
In this case the genes selected for the study are responsible for causing lung carcinoma and prostate carcinoma. Suffix P indicate prostate carcinoma and L indicates lung carcinoma. The resulting trees, as compared in earlier comparisons share 29.5 % topological similarity

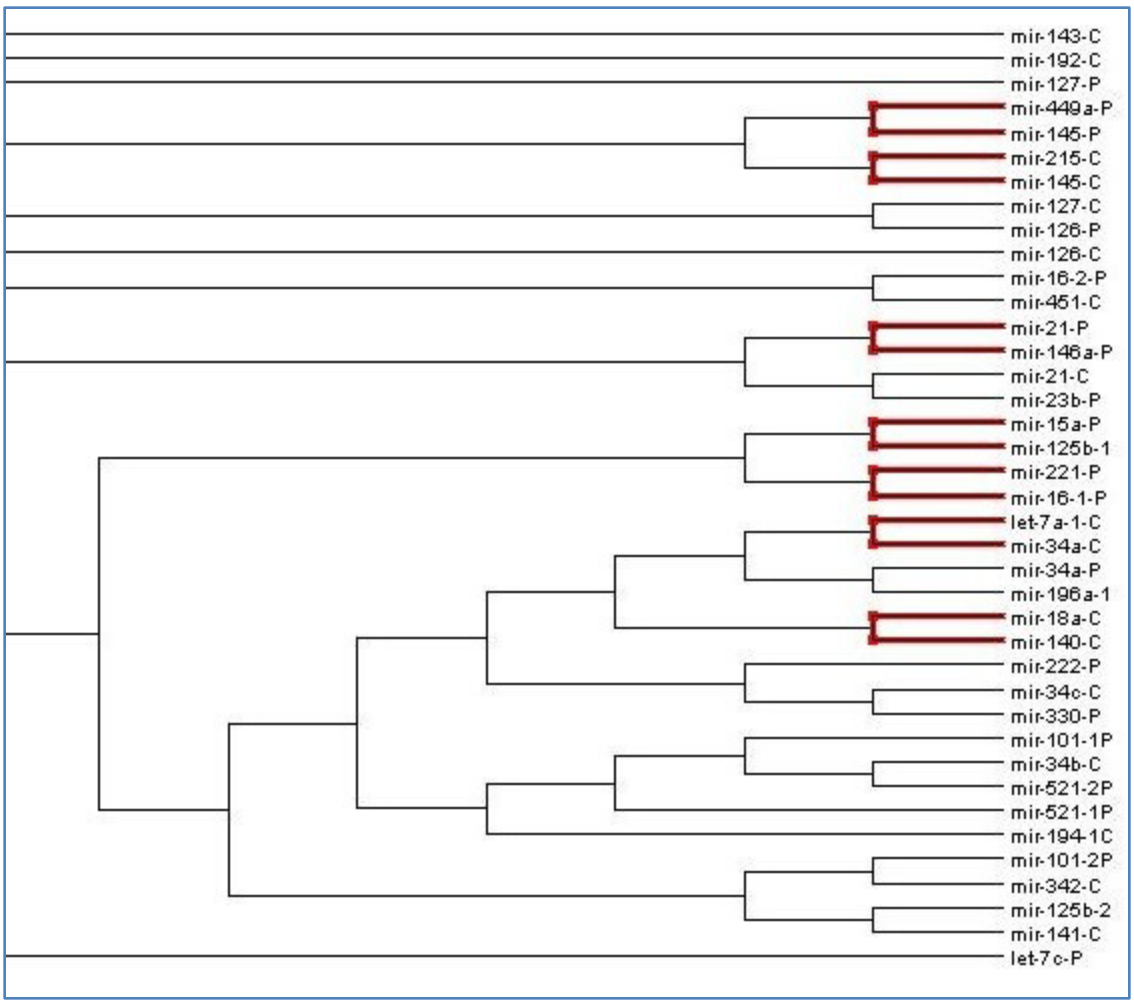


Figure 4. Phylogenetic tree showing the relationship between the colorectal cancer –Prostate cancers causing genes constructed using **sequence** data. Pair wherein both the genes have similar function is marked in red.

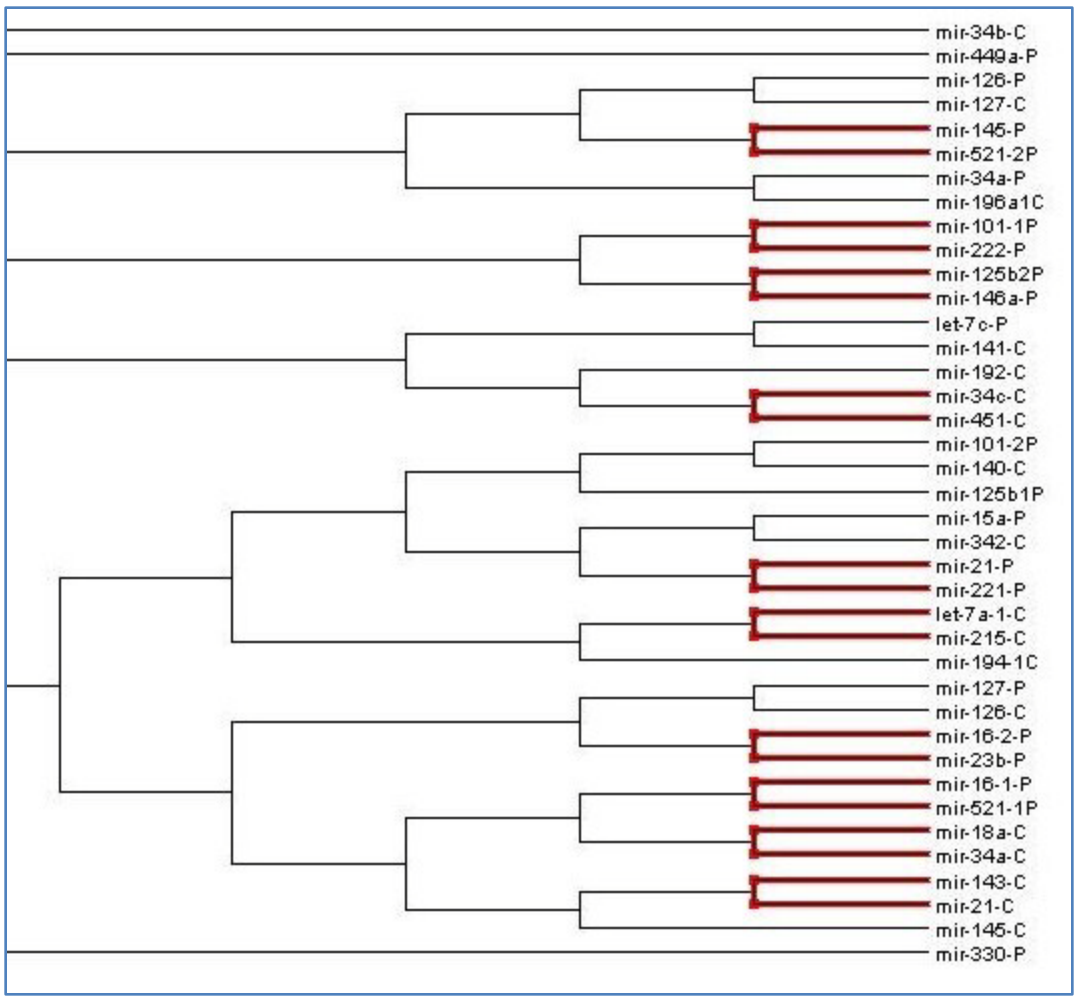


Figure 5. Phylogenetic tree showing the relationship between the Colorectal cancer –Prostate cancers causing genes constructed using **structure** data. Pair wherein both the genes have similar function is marked in red.

For the comparison let us consider a pair from the phylogenetic tree generated using structure data: has-mir-101-1 and has-mir-222. In the phylogenetic tree generated using the structure data these two genes are close, based on structural similarity, but in the phylogenetic tree generated from sequence data these genes are farther apart and are shown to be closer to has-mir-34b and has-mir-34c, respectively, which are the microRNA genes responsible for causing colorectal cancer. Whereas the gene pair we considered earlier is responsible for the causing prostate cancer. Similarly, there are many such genes which are closer in the phylogenetic tree generated using structure data and responsible for the same malfunction, but are shown to be more distantly related in the phylogenetic tree generated using sequence data.

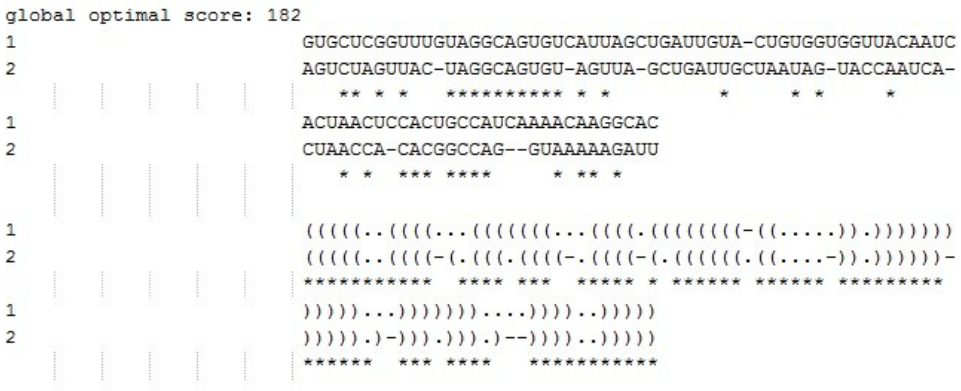
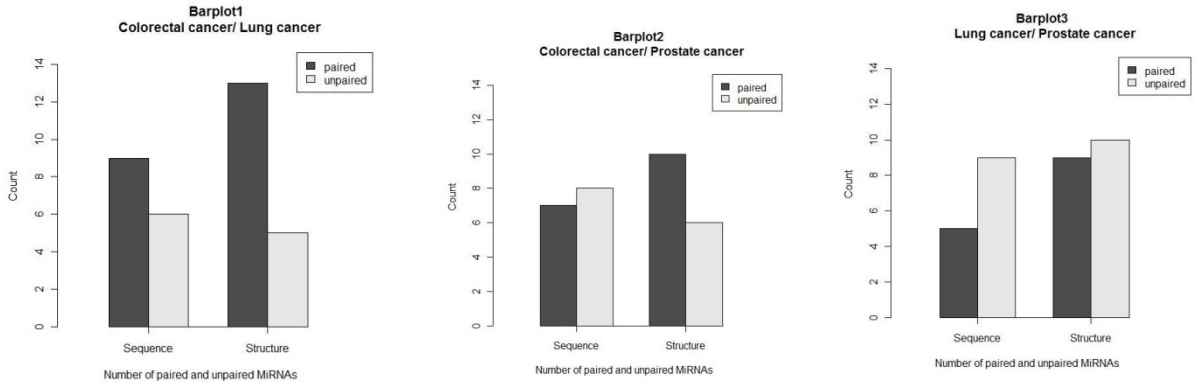


Figure 6. MicroRNA (has-mir-34b and has-mir-34c) shown in sequence based tree to be farther, if studied structurally are very similar and gives better similarity score.

When a clustering study is done on these genes we can see the genes which are shown to be farther in the sequence tree (has-mir-34b and has-mir-34c). if these genes are checked for their structural alignment they appears to be conserved in the alignment and also gives a better similarity score in RNAforester. In the phylogenetic tree based on secondary structure these genes are also farther in distance but are close to the genes responsible for causing colorectal cancer(has-mir-451).

The numbers of pairs causing the same malfunction and are close in the tree are counted for both types of trees, i.e., the tree generated using sequence data and structure data, respectively. Bar plots are plotted using these numbers (Figure 7A, 7B, 7C). By looking at the bar plot shown below we can say that the structure data give more number of such pairs than the sequence data.



(A) **(B)** **(C)**

Figure 7. (A) Bar plot showing the comparison of the phylogenetic data generated using microRNAs causing Colorectal cancer/Lung cancer; (B) Bar plot showing the comparison of the phylogenetic data generated using microRNAs causing Colorectal cancer/Prostate cancer; (C) Bar plot showing the comparison of the phylogenetic data generated using microRNAs causing Lung cancer/Prostate cancer.

These results are also compared statistically by applying Fisher's two sided exact test to the counted number of pairs (Fay, 2010). The test is implemented in R and the script is provided in Appendix B. Results of the Fisher's two sided exact test is given in Table 1. The test results shows that the alternative hypothesis is true and the phylogenetic tree based on structure information gives more number of functionally related pairs. But the confidence interval values used for the test are very high (see Table 1) (Fay, 2010) and thus we cannot completely rely on the statistical significance of the test values.

Table 1. Results of Fisher's two sided exact test.

STUDY	SELECTION	P-VALUE	CONFIDENCE VALUES
1	Colorectal/prostate	0.4875	2.6733
2	Colorectal/Lung	0.4795	2.5418
3	Prostate/Lung	0.7228	2.8397

Conclusion and Discussion

The first aspect in this study was to investigate whether there are differences in the phylogenetic relationships based on microRNA secondary structures and sequences, respectively. This study clearly showed that phylogenetic trees based on microRNA secondary structures have a different topology compare to sequence based phylogenetic tree.

A second aspect was to investigate if the functionally similar microRNAs are more closely related in phylogenetic tree based on secondary structure. Specifically here our goal is to investigate if microRNAs that are responsible for a particular disease or malfunction are more closely related to each other structurally than in their sequences.

To answer this question we counted the number of pairs from both type of the trees generated (i.e. tree generated by sequence data and the tree generated by structure data), such that both the paired genes are causing the similar type of cancer. A barplot of the data is obtained and we observe a trend wherein there are more number of functionally related gene pairs in the structure based trees than the sequence based trees.

Moreover, the analysis of phylogenetic trees based on structures can help in establishing a stable phylogenetic relationship, since secondary structures are more robust to the mutational changes.

This study also gives us an approach wherein we can establish a relationship between functionally related microRNAs, which can help to determine the function of newly discovered microRNAs. However, to show that this is achievable we need to extend the study by including more data and perform further analyses of this data. Although, it was observed that a trend exist for more number of functionally related pairs in the structure based study than in the sequence based study. For example in figure 4 we can see there are 7 gene pairs whose function is similar whereas in figure 5 we can see there are 10 such pairs.

Future Work

This current study is mostly done to investigate the relationship between the phylogenetic tree constructed using microRNA sequence and structure. In order to further investigate functional relation of microRNAs, those that cause colorectal cancer, lung cancer and prostate cancer were also studied in a phylogenetic context. A few more studies involving miRNAs causing leukemia, stomach cancer can be performed. To generalize this approach a web server could be set up that produces the phylogenetic relationship of newly discovered microRNA and those whose function is well established. Currently there are approaches that use experimentally verified miRNA pathways (Ulitsky et al., 2010) or approaches that use target prediction in combination with expression profiling to predict the microRNA function (Wang, 2006). However, our approach mainly focuses on the secondary structure similarity in phylogenetic context and thus can show a more robust relationship between the microRNAs and therefore complement previous approaches.

References

ALTUVIA, Y., LANDGRAF, P., LITHWICK, G., ELEFANT, N., PFEFFER, S., ARAVIN, A., BROWNSTEIN, M. J., TUSCHL, T. & MARGALIT, H. 2005. Clustering and conservation patterns of human microRNAs. *Nucleic Acids Res,* 33**,** 2697-706.

ALVAREZ-GARCIA, I. & MISKA, E. A. 2005. MicroRNA functions in animal development and human disease. *Development,* 132**,** 4653-62.

BASSI, S. A. G., VIRGINIA. 2007. *DNALinux Virtual Desktop Edition. Available from Nature Precedings* [Online]. Available: http://www.dnalinux.com/ [Accessed].

BEREZIKOV, E., GURYEV, V., VAN DE BELT, J., WIENHOLDS, E., PLASTERK, R. H. & CUPPEN, E. 2005. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell,* 120**,** 21-4.

CARVER, T. 2010. *Distmat* [Online]. Available: http://emboss.open-bio.org/wiki/Appdoc:Distmat#Author.28s.29 [Accessed may 15 2010].

CONGDON, C. B. 2006. An Evolutionary Algorithms Approach to Phylogenetic Tree Construction. *Springer London***,** 99-116.

FAY, M. P. 2010. Exact Conditional Tests and Matching Confidence Intervals for 2 by 2 Tables.

FELSENSTEIN, J. 1995. PHYLIP (Phylogeny Inference Package) Version 3.57c.
FRIEDMAN, R. C., FARH, K. K., BURGE, C. B. & BARTEL, D. P. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res,* 19**,** 92-105.
GRIFFITHS-JONES, S. 2006. miRBase: the microRNA sequence database. *Methods Mol Biol,* 342**,** 129-38.
HOCHSMANN, M., VOSS, B. & GIEGERICH, R. 2004. Pure multiple RNA secondary structure alignments: a progressive profile approach. *IEEE/ACM Trans Comput Biol Bioinform,* 1**,** 53-62.
HUSON, D. H., RICHTER, D. C., RAUSCH, C., DEZULIAN, T., FRANZ, M. & RUPP, R. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics,* 8**,** 460.
JIANG, Q., WANG, Y., HAO, Y., JUAN, L., TENG, M., ZHANG, X., LI, M., WANG, G. & LIU, Y. 2009. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res,* 37**,** D98-104.
KJER, K. M. 1995. Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs. *Mol Phylogenet Evol,* 4**,** 314-30.
LIANG, H. & LANDWEBER, L. F. 2005. Molecular mimicry: quantitative methods to study structural similarity between protein and RNA. *RNA,* 11**,** 1167-72.
LU, R. C. T. L. C. L. 2010. *RNA Secondary Structure Prediction* [Online]. Algorithms for Molecular Biology. Available: http://163.22.21.49/course/biology/slidef2_rna.pdf [Accessed].
MATTHIAS HÖCHSMANN, T. T., ROBERT GIEGERICH, STEFAN KURTZ: 2003. Local Similarity in RNA Secondary Structures. *IEEE Bioinformatics Conference 2003***,** 159-168.
MCCASKILL, J. S. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers,* 29**,** 1105-19.
NYE, T. M., LIO, P. & GILKS, W. R. 2006. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics,* 22**,** 117-9.
PRIMER, S. 2004. *SYSTEMATICS AND MOLECULAR PHYLOGENETICS* [Online]. NCBI Science Primer. [Accessed 2010].
STAPLE, D. W. & BUTCHER, S. E. 2005. Pseudoknots: RNA structures with diverse functions. *PLoS Biol,* 3**,** e213.
THOMPSON, J. D., HIGGINS, D. G. & GIBSON, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res,* 22**,** 4673-80.
ULITSKY, I., LAURENT, L. C. & SHAMIR, R. 2010. Towards computational prediction of microRNA function and activity. *Nucleic Acids Res,* 38**,** e160.
WANG, X. 2006. Systematic identification of microRNA functions by combining target prediction and expression profiling. *Nucleic Acids Res,* 34**,** 1646-52.
WANG, X., ZHANG, J., LI, F., GU, J., HE, T., ZHANG, X. & LI, Y. 2005. MicroRNA identification based on sequence and structure alignment. *Bioinformatics,* 21**,** 3610-4.
ZUKER, M. & STIEGLER, P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res,* 9**,** 133-48.

Appendix A: Phylogenetic trees showing the results of the study 2 and study 3.

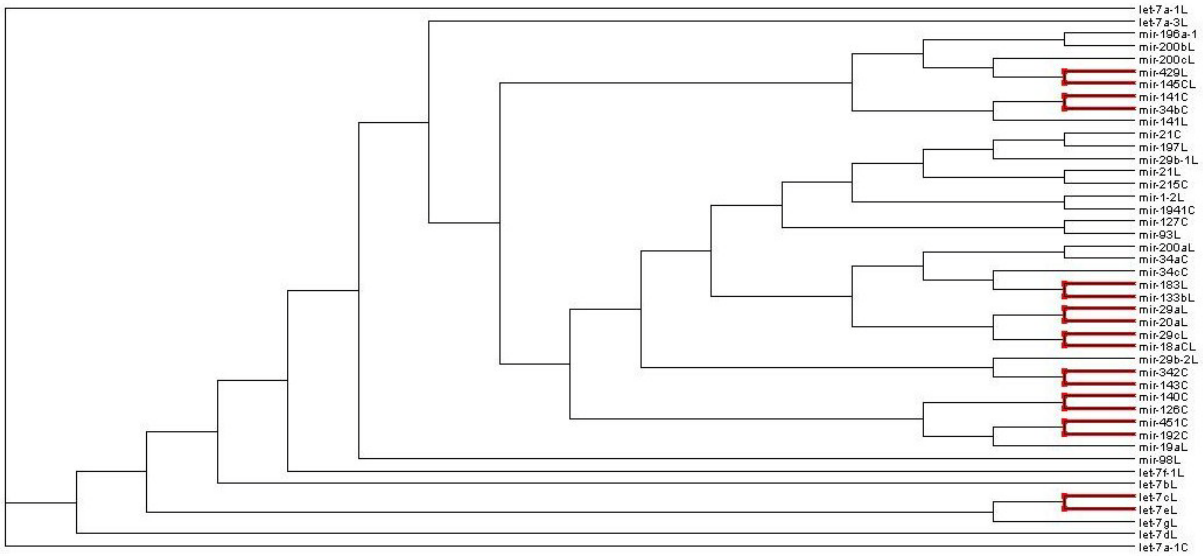


Figure 8. Phylogenetic tree showing the relationship between the colorectal cancer –Lung cancers causing genes. Constructed using **sequence** data

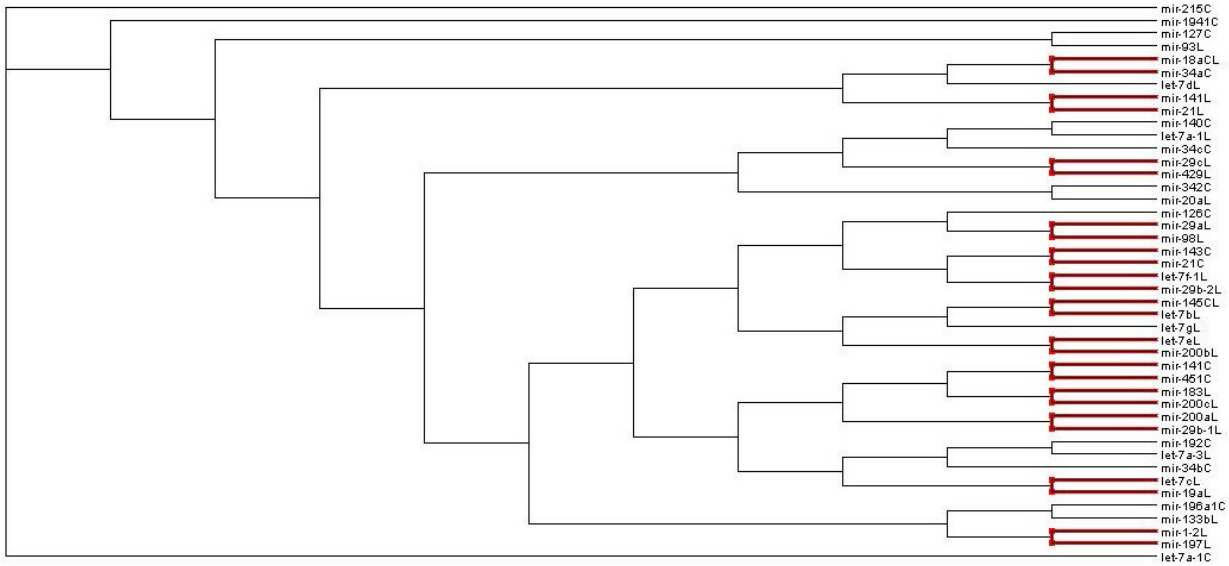


Figure 9. Phylogenetic tree showing the relationship between the colorectal cancer –Lung cancers causing genes. Constructed using **structure** data

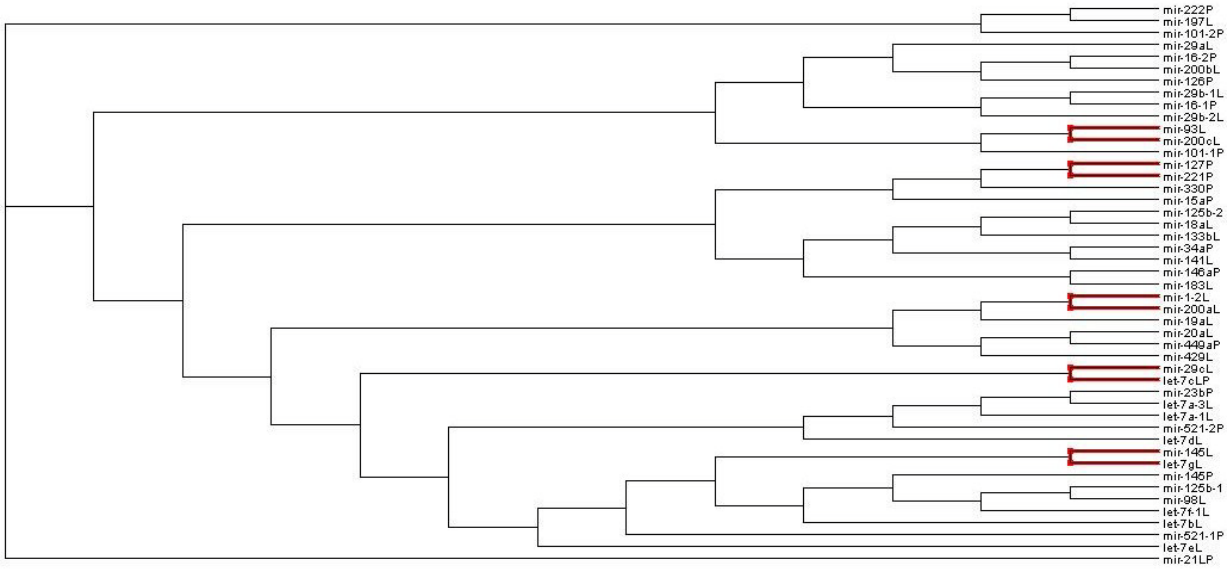


Figure 10. Phylogenetic tree showing the relationship between the colorectal cancer –Lung cancers causing genes. Constructed using **structure** data

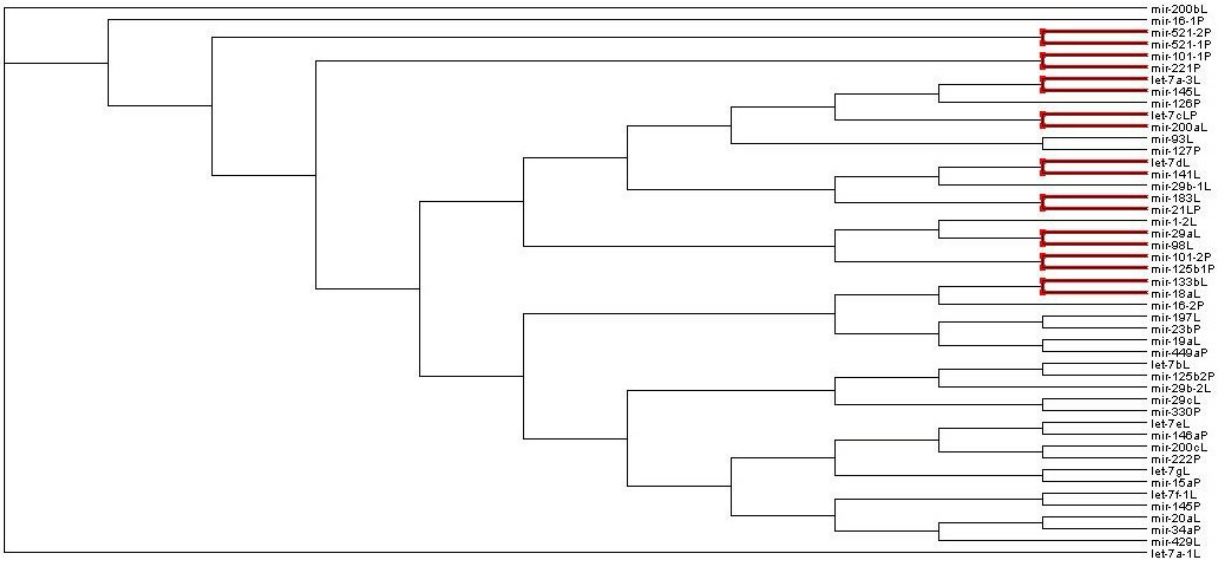


Figure 11. Phylogenetic tree showing the relationship between the Lung cancers – Prostate cancer causing genes. Constructed using **structure** data

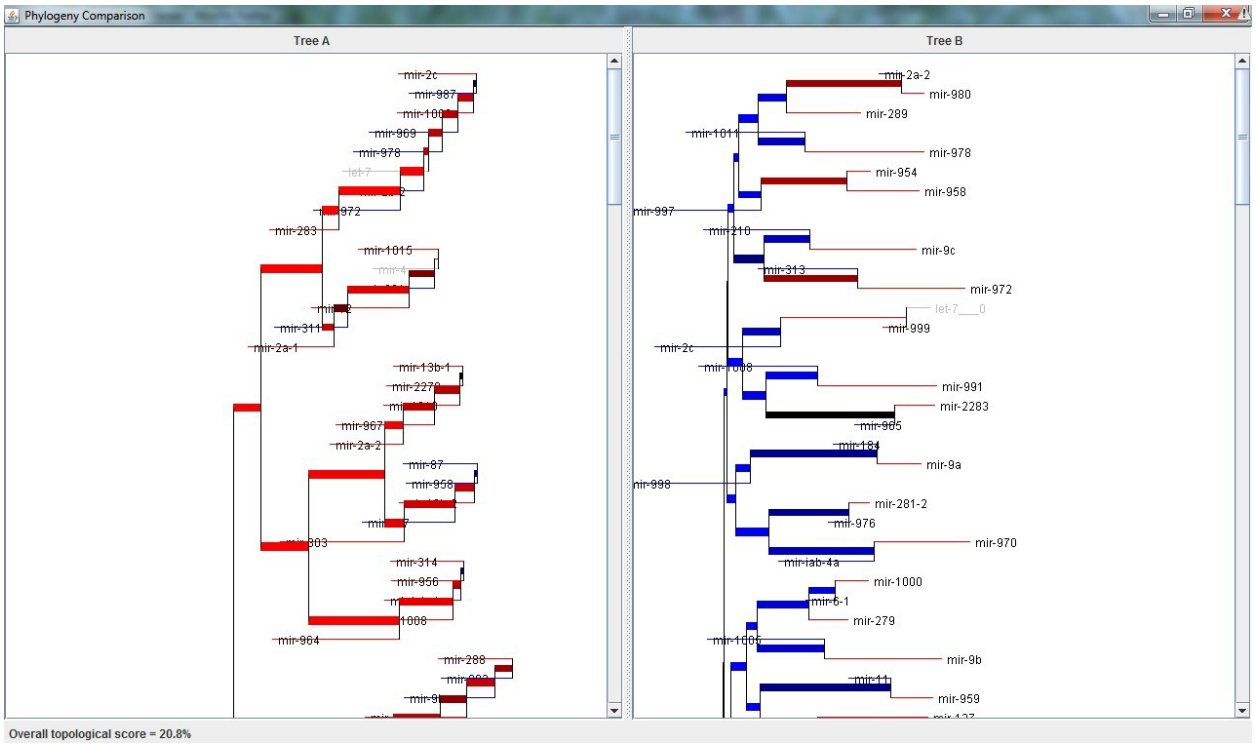


Figure 12. Comparison of the topology of the two phylogenetic trees generated for the organism *Drosophila melanogaster*. On the left side there is tree generated using sequence data and on the right side there is tree generated using structure data.

Appendix B: legend for R code

R programming script for the fisher's exact test.

```
ft <- fisher.test(x)
```

```
x <- matrix(c(9, 13, 6, 5), 2, 2, dimnames = list(c("paired", "unpaired"),
c("Sequence", "Structure"))) # assign a variable for the contingency table 1

y <- matrix(c(7, 10, 8, 6), 2, 2, dimnames = list(c("paired", "unpaired"),
c("Sequence", "Structure"))) # assign a variable for the contingency table 2

z <- matrix(c(5, 9, 9, 10), 2, 2, dimnames = list(c("paired", "unpaired"),
c("Sequence", "Structure"))) # assign a variable for the contingency table 3

      exact2x2(x, tsmethod = "minlike")

      exact2x2(y, tsmethod = "minlike")

      exact2x2(z, tsmethod = "minlike")
```