Determining the Evolution of Cancer: A Case Study in Phylogenetic Analysis

Kymberleigh Pagel Institute: Texas A&M University, Indiana University email:kpagel@indiana.edu
Suzanne Matthews Institute: Texas A&M University email:sjm@cse.tamu.edu
Tiffani L Williams Institute: Texas A&M University email:tlw@cse.tamu.edu

***Abstract*—Phylogenetic tools are useful in determining the evolutionary relationships between a group of organisms. In this study, we trace the evolutionary history of cancer by using phylogenetic analysis. We selected and analyzed a set of nine genes across eight organisms for both amino acid and nucleotide sequences. By getting a better understanding of how these genes change over time, we can determine the mechanisms of mutation in these genes and gain new insights on how cancer is formed. Multiple sequence alignment was performed for both nucleotide and protein sequences for all genes. We performed exhaustive tree search on both individual genes and concatenated sequences to generate the set of all trees and the maximum likelihood trees. Using a number of post-processing and visualization tools, we were able to study how the phylogenetic search proceeded through tree space, and determine a likely hypothesis for the evolution of these cancer genes.**

***Index Terms*—Phylogeny, Cancer, Tree Searching, Evolution**

I. Introduction

A. What is Phylogeny?

Phylogenetics is the study of evolutionary relationships between organisms. These relationships can be estimated by finding morphological and genetic similarities. By performing phylogenetic analysis across multiple organisms, a tree can generated. This tree represents a hypothesis of the evolutionary relationship between these organisms. Phylogenetic analysis allows for the clear understanding of the similarities and differences between species. The understanding of the relationship between organisms sets the framework for many other fields of research. Phylogeny is especially useful in studying the development of diseases. Determining the pattern of evolution of sensitive genes across species can help to pinpoint the critical differences and explain the mechanisms of the disease. Phylogeny is of critical importance in our understanding of the evolution of species in the past and into the future.

B. Evolution and Cancer

The human genome is composed of several thousand genes, each of which code for at least one protein. Proteins are the molecular machines which carry out the work of the body and are essential to sustain our lives. Over time however the genome can be host to mutations which can affect the coding regions of the DNA. The vast majority of these mutations will have a negative effect. An acculmulation of deleterious mutations, insertions, deletions, and transformations can lead to the formation of cancer, especially if the mutations occur in genes which code for pivotal proteins. The genes we have chosen in this study have been linked to the formation of cancer because they perform functions related to cell differentiaion and cell cycle regulation.

An example of a cancer caused by disrupted cell cycle regulation is myelotic leukemia. In this form of leukemia the body overproduces white blood cells. The cells are produced so quickly and in such an abundance that there is not enough time for the cells to mature. Accumulation of the abnormally formed white cells interferes with the production of normal red and white blood cells. One of the genes covered in this study, Dap5, is thought to be associated with leukemia.

C. Overview of Paper

In this study we use phylogenetic analysis to create a hypothesis tree which approximates the evolution of a set of eight genes with close ties to cancer. In Section II we summarize the process of phylogenetic analysis and provide brief summaries of the programs we used to perform our analysis. In Section III we describe the data used to perform our analysis specifying the genes and model organisms included in this study and why they were chosen. Section IV provides specific information on how we performed our phylogenetic analysis. Lastly we review the results of our analysis and describes the meaning behind our results in Section V.

II. Overview of Phylogenetic Analysis

A. Multiple Sequence Alignment

The first step of phylogenetic analysis is to highlight the similarities between gene sequences. By comparing gene sequences, it is possible to infer evolutionary relationships. Organisms which are more closely related should have more similar sequences, while distantly related organisms may only share a very small portion of the gene. Multiple sequence alignment is a technique to identify regions of the sequence that are conserved between several organisms. Sequence alignment is the process of arranging the sequences such that very conserved sections of the sequence are directly compared and highlighted. These aligned segments are seperated by gaps such that the maximum amount of the sequence is aligned between the organisms. The use of gaps also serves to highlight the regions of code which are dissimilar. Regions of gene sequences which are conserved between many organisms are most likely to serve some important function. Differences in sequence could be caused by point mutations and the long

gaps can be caused by insertions/deletions. The difficulty of sequence alignment is compounded when there are extensive insertions, deletions, point mutations, and most importantly when there are more than two sequences to be compared.

B. Tree Searching

Tree searching is a process that uses a sequence alignment to create and analyzes all possible trees which represent possible evolutionary relationships between a set of organisms. The trees are compared to see which the most likely to represent the true evolutionary relationships. The goal of tree searching is to highlight the tree which represents the most likely evolutionary relationship between the organisms.

1. Models of Evolution
   One of the parameters which needs to be determined before the differences are analyzed is the model of evolution. A model of evolution, also called a substitution model, describes the way that the gene sequence changes over time. The model of evolution specifies the approximate frequencies for nucleotides or amino acids and also the rate at which they change. In our study, we are looking at model organisms which are distantly related so specifying a model of evolution is very important. The use of a model during tree searching allows for a deeper understanding of the sequence changes between organisms.

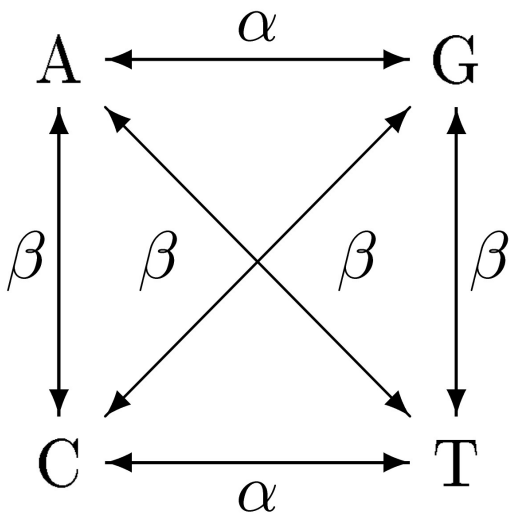


Figure 1. The two-parameter model of evolution proposed by Kimura (k2p) [10].

An example of an evoutionary model is the two-parameter Kimura model [6]. This model of evolution is for nucleotides, rather than amino acids. The model, illustrated in Figure 1, sets substitution rates to be equal except in the cases of transitions and transversions. The horizantal arrows represent a transitions which occur at a rate equal to α, the other arrows occur at a rate β.

2. Tree Searching
   Tree searching is the creation and analysis of possible trees to select the trees with the best score. There is considerable difficulty in generating and scoring all trees when the size of tree space is very large. The size of tree space is equal to $(2n\text{-}5)!!$ where n is the number of taxa. If there are more than fifteen organisms then it is not feasible to perform an exhaustive search which evaluates the score for every possible topology, because there are billions trees in tree space. In our study we evaluate (n=8) eight organisms resulting in 10,395 possible trees to traverse in tree space. This allows us to easily perform an exhaustive tree search which enumerated and scored every possible tree. We also performed maximum likelihood tree searches, which searched through every possible topology and resulted in the selection of the tree with the highest likelihood score.

C. Post-Processing & Visualization

Exhaustive tree searching on a small space can generate tens of thousands of trees, making it impossible for a person to do comparison and analysis by hand. At the same time, it is necessary to compare this many trees because we want to know how the search proceeded through tree space. To compensate for the difficulties associated with the abundance of trees, we use post-processing and visualization techniques to select the most likely hypothesis. Post-processing allows us to reduce the number of trees that need to be compared by eliminating extraneous data and reducing the large data set down to representative samples. The reduced data set is then analyzed using visualization tools. Visualization is key to comparison and understanding results of tree searching. In our study we used visualization tools such as heatmaps, multidimensional scaling, and tree plotting to compare the different tree searches.

1. Comparing trees
   While not every one of the thousands of the trees can be compared by hand, we can select a representative sample that will provide equivalent results.
   The comparison of individual trees allows for analysis not only on the level of important partitions but also trends present in the majority of trees. A simple analysis would be to compare the frequency of a common partition among several trees. For example the partition which contains two key organisms that are very closely related should be present in the majority of trees. If this is not the case then there might be significant errors or the organisms might not be as closely related as we expected. To generate the tree comparison we would look at the full set of trees we have to compare, count up the number of trees which have this partition and then divide that by the total number of trees which include both organisms.
2. Heatmaps

A classic way to visualize large sets of data is to use a heatmap. Rather than looking at a page full of numbers, a heatmap allows the viewer to recognize patterns at a glance. A heatmap is a graphical representation of a set of data which displays the data using colors. In our plots higher values are represented by darker colors whereas smaller values are denoted with lighter colors.

3. Multidimensional scaling
   Another way of visualizing trees is to use multidimensional scaling. In this method each tree is represented as a point in two dimensional space. The distance between points on the plot corresponds to the distance between the trees. This is useful for comparing the trees generated from one or more sets of trees generated using tree searches with different sequence data or substitution models.

Post-processing and visualization are key to understanding the meaningful differences between sets of trees. Using multidimensional scaling and heatmaps we can look at the general trends present in the data. By comparing trees directly we can not only evaluate overall trends in the topologies but also highlight key bipartitions and perform comparison of the smaller details.

III. Data

We collected the amino acid and nucleotide sequence data for eight organisms and eight genes which were extracted from the HomoloGene [7] database. In order to maintain a high degree of sequence validity, we selected sequence data which annotated with the RefSeq status of reviewed, validated or model. The inclusion of these statuses means that the data we used in our study was could be assumed to be accurate rather than generated by a computer based upon homology. There was not equal coverage for all genes on all organisms. The coverage for genes and organisms is displayed in the Table 1.

	arc	cmyc	cyclind	dap5	lmyc	oct4	p27	xiap
Human	1	1	1	1	1	1	1	1
Chimp	1	1	0	1	1	1	1	1
Dog	0	1	0	1	1	1	0	1
Cow	1	1	0	1	1	0	0	1
Mouse	1	1	1	1	0	1	1	1
Zebrafish	0	1	1	1	0	0	0	0
Fruit fly	0	0	1	1	0	0	0	0
Chicken	0	0	0	1	1	0	1	0

Table 1. Sequence Availability. A value of one represents a sequence which is present, a value of zero, a sequence which is absent.

A. Genes Under Study

The genes we studied were selected due to their implications in cancer. We briefly describe below.

ARC Full name: apoptosis repressor with a CARD domain. Serves to inhibit programmed cell death, also called *apoptosis*, in muscle cells. This protein is expressed in high levels in cancer cell lines.

C-myc Full name: MYC. Helps to regulate cell cycle progression, apoptosis and cellular transformation. Functions as a transcription factor that regulates transcription of specific target genes. A mutation which results in persistent expression of the gene can results in the formation of cancer.

CyclinD1 Full name: G1/S-specific cyclin-D1. Regulates cell cycle progression. Overexpression or amplification of this gene can contribute to tumorgenesis.

Dap5 Full name: Death Associated Protein 5. Inhibits translational initiation, differentiation, and apoptosis, associated with leukemia.

Lmyc Full name: L-myc-1 proto-oncogene protein. Believed to participate in the control of cell proliferation and differentiation. If mutated can be involved in tumorgenesis.

Oct4 Full name: Octamer 4. Promotes cell differentiation. Mutation causes an inhibition of cellular differentiation can result in cancer formation in adult germ cells.

p27 Full name: Cyclin-dependent kinase inhibitor 1B. Regulates cell cycle progression. Mutation can disrupt cell cycle regulation and lead to uncontrolled cellular proliferation.

Xiap Full name: X-linked Inhibitor of Apoptosis Protein. Inhibits apoptosis. Overexpression of this gene can result in cancer, autoimmunity, and neurodegenerative disorders.

B. Organisms Under Study

The organisms which were selected are model organisms which have varying degrees of cancer incidence but all show promise in their ability to help us better understand cancer. The organisms are:

Homo sapiens Common name: human. Humans have an abnormally high rate of cancer.

Pan troglodytes Common name: chimpanzee. Despite the high genetic similarity to *Homo sapiens*, chimps have low levels of cancer. Close comparison of the genes linked to cancer is needed to fully understand this phenomenon.

Canis familiaris Common name: dog. The high rate of cancer in this organism is due to the fact that it tends to life into old age far more often than *Pan troglodytes* or *Bos taurus*.

Bos taurus Common name: bovine/cow. A good model organism for comparison against primate and rodent models.

Mus musculus Common name: mouse. Have a high incidence of cancer. This is a good model organism due to its small size and short lifespan.

Danio rerio Common name: zebrafish. High incidence of cancer but this organism can be induced to host a broad spectrum of human cancers [2].

Drosophila melanogaster Common name: fruit fly. Low incidence of cancer. This is a very good model organism because of its very small size and short life span. Although the organism is very different from *Homo sapiens* physiologically, many of the cellular processes are very similar.

Gallus gallus Common name: chicken. Moderate to low incidence of cancer. An excellent model organism for study due to the easy availability of embryos.

IV. Methods

A. Multiple Sequence Alignment

Multiple sequence alignment was performed using WebPrank [11]. We generated alignments for both nucleotide and amino acid sequences on individual genes and for the concatenated sequences. For the nucleotide alignment we chose the option to align translated proteins rather than to align the nucleotides, in order to increase accuracy. We selected the options to trust insertions and compute reliability.

B. Tree Searching

To perform the exhaustive maximum likelihood tree searching with nucleotide sequences I used a program called PAUP* [17]. PAUP*, which stands for Phylogenetic Analysis Using Parsimony, is a tool which accepts an input of a sequence alignment and returns a set of trees. The program operates by using code blocks to specify commands.

We performed tree searches using PAUP* on the nucleotide sequences for every gene and for the concatenated gene sequences. For each sequence, we performed tree searching to produce both the set of all possible trees and the single maximum likelihood tree. The first search was an exhaustive maximum likelihood search which sought to create the full set of 10,395 trees. The second was a maximum likelihood search which aimed to produce the single tree with the highest score. Additionally, we performed these tree searches six times per gene for each substitution model available on PAUP*. The substition models available for use with PAUP* are: Jukes and Cantor (JC69) [9], Felsenstein (F81) [4], Kimura (K2P) [10], Felsenstein (F84) [5], General Time-Reversible (GTR) [16], and Hasegawa, Kishino and Yano (HKY) [8].

To ensure the search was exhaustive we utilized the command `alltrees` with a `keep` value of 1,000,000 and increased the `maxtrees` to 11,000. To create a single maximum likelihood tree we reduced the value of `keep` down such that PAUP* only retained the best scoring trees.

We were able to produce the full set of trees for both f84 and hky (1985), but the other sets of trees could not be created due to unavoidable restrictions within the program.

PAUP* is not capable at performing tree searching on amino acid sequences using the maximum likelihood criterion. Instead, we chose a program called PhyML [15]. PhyML is similar to PAUP* except that it accepts commands in a more standard command line interface instead of code blocks. Prior to performing tree searching we utilized a program called ProtTest [1] to determine the most likely substituion model for the protein sequences; the model Jones Taylor Thornton (JTT) [3] was selected. It was necessary to reduce the set down to one model due to the fact that it took nearly a day to perform tree searching on one gene.

C. Post-Processing

To assist in post-processing and analysis of our collection of trees we used an algorithm called Phlash [12], a fast algorithm for comparing trees via the creation of similarity or distance matrices. Phlash accepts an input of two groups of trees and then uses the distance between each set of two trees to creates a similarity matrix. Rather than directly compare each possible tree visually, it is important that we consider the relationship between the trees generated. To this end we used Phlash to create the similarity matrices for each set of trees and to compare the results from the two substition models within the nucleotide sequences. The full similarity matrices were created using RF Rate as the distance measure. RF rate [12] is the normalization of the Robinson-Foulds distance, calculated by dividing the RF distance by the number of non-trivial bipartitions in the tree. The Robinson-Foulds [14] metric is a method for calculating the distance between unrooted trees. We are using this metric to compare trees generated in tree searching.

V. Results & Discussion

We created several multidimensional scaling plots and a heatmap to compare the trees created within and between the full sets of trees. Although we generated a total of 10,395 trees for each search of tree space, we only included a set of 1042 trees in the visualization to reduce the size of the image and make the plots easier to comprehend. To generate the set of 1042 trees, we created a new list composed of every tenth tree from the original set of 10,395. Although the data set was reduced, the distribution of the trees being compared is still approximately equivalent to the full set.

Using multidimensional scaling plots, we compared the trees created within the tree searching. Each point on the plot represents a tree; the distance between the points represents the distance between the trees. We created multidimensional scaling plots for the trees generated using different substitution models to see how the runs of tree searching traversed tree space differently.

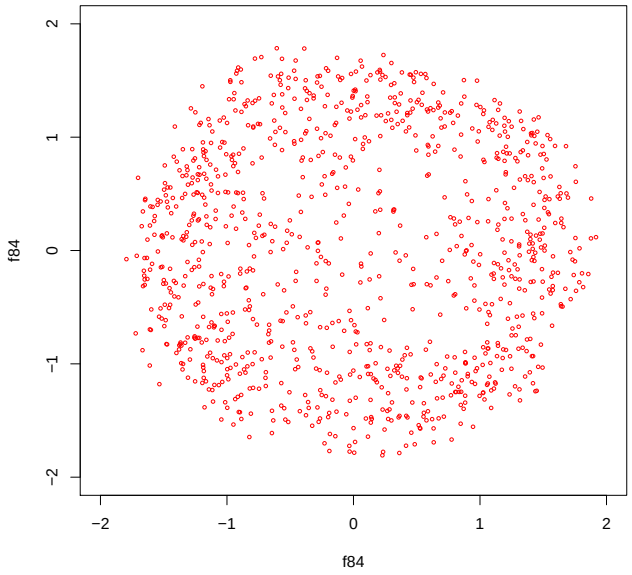


Figure 2. An MDS plot which displays 1042 of the trees created with tree searching using PAUP* for the concatenated sequence with the F84 substitution model.

The positions of the points in Figure 2 show that there are some trees generated in the search of tree space which are very similar, indicated by the darker clustering of points. In general the search seemed to have produced a fairly diverse set of trees which are not very similar to one another. The plot generated for tree searching using the hky substition model showed a similar distribution.

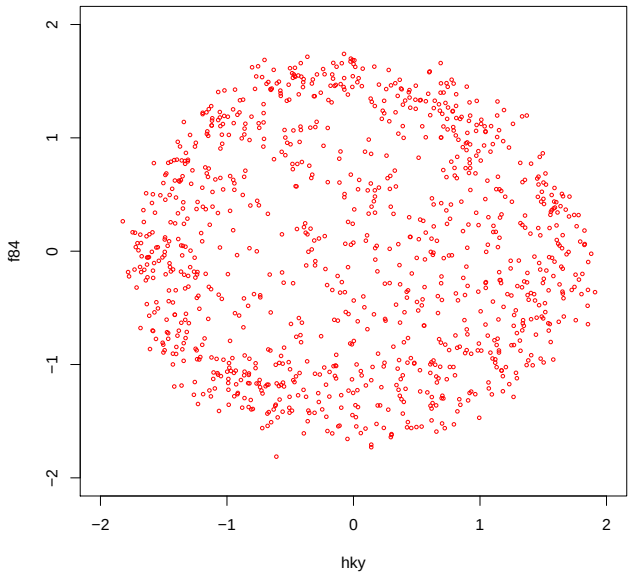


Figure 3. An MDS plot which displays 1042 of the trees generated with tree searching on the concatenated nucleotide sequence for the substitution models F84 and HKY. The selection of the hypothesis tree was based upon the similarities between the trees generated by the f84 and hky substitution models.

This plot has a distribution comparable to those produced by the individual substitution models. The MDS plots show that both of those tree searches proceeded through tree space in a similar manner. This pattern could be in part due to the fact that the two models are much alike. Both F84 and HKY allow for unequal base frequencies, but there is a difference in how the models handle the transition/transversion ratio. Although the two substitution models have much in common the difference in ratios could explain why the two tree seaches traversed tree space slightly differently.

The similarities between the trees produced by the two different substition models lends support to the tree generated with the concatenated nucleotide sequence as a likely hypothesis.

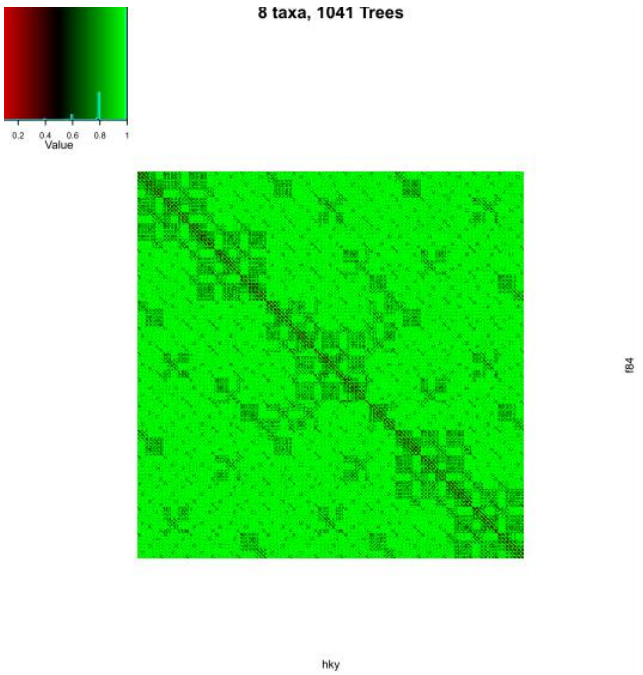


Figure 4. A heatmap which displays the similarity matrix generated in Phlash comparing the 1042 trees between the two substitution models.

Additionally we created a heatmap to compare the results of tree searching when using the models F84 and HKY. We compared the sets of 1042 trees, rather than the full set of 10,395. The heatmap was created using the `heatmap2` function within the gplots package in R. The light green color prominent in the image means that the trees generated with the different models are very dissimilar.

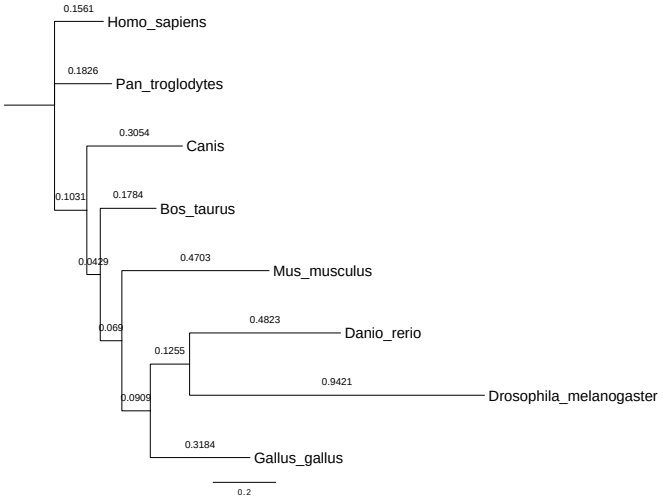
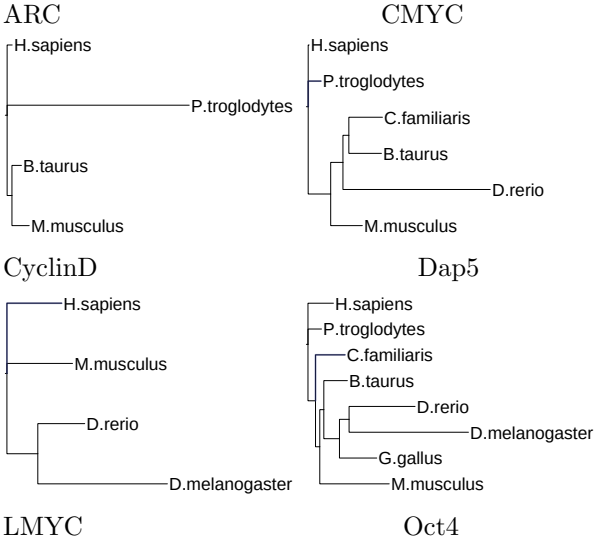


Figure 5. FigTree [13] was used to visualize the maximum likelihood tree for the concatenated nucleotide sequence. Although there were differences in branch lengths, the topology of the trees were exactly the same.

We have selected the maximum likelihood tree generated by the tree searching for the concatenated nucleotide data as the likely hypothesis. This tree shows the evolutionary relationships of the cancer genes in the selected organisms. The main disparity between the hypothesis trees generated for the concatenated amino acid and nucleotide sequences was the position of chicken relative to mouse. If we were to remove mouse from the trees then the layout would be the same.

There were only sequence data available for three genes of chicken. In future studies it would be beneficial to include genes and organisms which are very well covered. Additional comparisons should be made with the different substitution models, to ensure that the other models do not produce very different models.



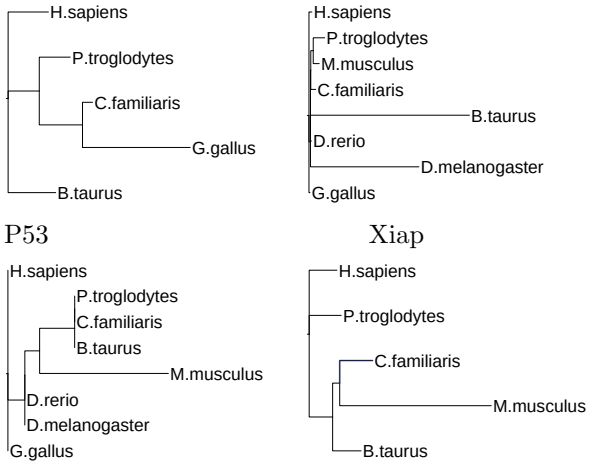


Figure 6. The eight gene trees for f84 substition model. Only eight of the sixteen trees are displayed because the substition models produced trees with identical topologies.

Additionally we created the maximum likelihood tree for each gene and evaluated the most common bipartitions. The trees generated for each individual genes, displayed in Figure 6, looked very different from one another. The trees did not have the same amount of partitions, seven out of the eight gene trees contained less than the full set of organisms due to uneven coverage. The topologies generated for each gene differed not only on the numbers of bipartitions but also the relative positioning of the organisms. In the arc tree mouse and cow are sister taxa, while in the xiap tree mouse and dog are sister taxa. This is in part due to the uneven coverage of the genes. A tree which corresponds to the gene arc only has four organisms while xiap has five organisms. The relative frequency of the most common bipartitions are displayed in the table below:

organisms	nucleotide trees	aa trees
human, chimp	6/7	5/7
human, chimp, dog	3/3	1/6
human, chimp, dog, cow	0/5	1/5
fruit fly, zebrafish	1/2	3/3
fruit fly, zebrafish, chicken	0/2	0/2

Table 2. Frequencies of key organism groups in the trees generated with individual gene data.

The first column of Table 2 consists of groupings of organisms which are present in the concatenated sequence tree. The second column shows the number of individual gene trees which contain that bipartition divided by the total number of trees which contain the specified organisms. The third column is the same as the second, except for gene trees generated with the amino acid sequences. For example, the first row of Table 2 indicates that the grouping of human and chimp occurs six times in the individual nucleotide gene trees, but both organisms are present in seven out of the eight trees. This means that one tree contained both human and chimp yet the organisms were not on the same branch.

Table 2 indicates that although the topologies for the gene trees are highly different, there are persistent similarities. These common branches between the trees can manifest themselves in the tree generated from the concatenated sequence. For both amino acid and nucleotide sequences there is a trend which indicates that bipartitions present in the majority of the individual gene trees tend to be present in the concatenated sequence tree. This lends more credibility to the selection of the concatenated nucleotide sequence tree as a likely hypothesis.

References

[1] F. Abascal, R. Zardoya, and D. Posada. Prottest: Selection of best-fit models of protein evolution. *Bioinformatics*, 21(9):2104–2105, 2005.

[2] J. F. Amatruda, J. L. Shepard, H. M. Stern, and L. I. Zon. Zebrafish as a cancer model system. *Cancer Cell*, 1(3):229–231, 2002.

[3] J. DT, T. WR, and T. JM. The rapid generation of mutation data matrices from protein sequences. *Comput Applic Biosci*, 8:275–282, 1992.

[4] J. Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[5] J. Felsenstein. Distance methods for inferring phylogenies: A justification. *Society for the Study of Evolution*, 38(1):16–24, 1984.

[6] J. Felsenstein. *Inferring Phylogenies*. Sinauer, 2004.

[7] M.-B. A. G. R. H. L. H. J. H. S. L. C. S. W. B. S. Geer LY, A. Marchler-Bauer, G. RC, L. Han, J. He, S. He, S. W. Liu C, W. Shi, and B. SH. The ncbi biosystems database. *Nucleic Acids Res.*, 38:492–6, 2010.

[8] M. Hasegawa, T. aki Yano, and H. Kishino. Dating of the human-ape splitting by a molecular clock. *Journal of Molecular Evolution*, 22(2):160–174, 1985.

[9] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. *Mammalian protein metabolism*, 3:21–132, 1969.

[10] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120, 1980.

[11] A. Löytynoja and N. Goldman. webprank: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics*, 11(579), 2010.

[12] S. Matthews, M. L. Smith, and T. L. Williams. The abcs of phylogenetic comparison. 2011.

[13] A. Rambaut. Figtree, a graphical viewer of phylogenetic trees, 2007.

[14] D. F. Robinson and L. R. Foulds. Comparison of weighted labelled trees. *Combinatorial mathematics*, 748, 119-126 1979.

[15] G. S., D. J.F., L. V., A. M., H. W., and G. O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of phyml 3.0. *Systematic Biology*, 59(3):307–21, 2010.

[16] T. S. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.

[17] D. L. Swofford. Paup*. phylogenetic analysis using parsimony (*and other methods). version 4. Sinauer Associates, Sunderland, Massachusetts., 2003.