Application of Graph Theory in Mathematical Biology

MAT 4011 Capstone
Ashley Holliday

Abstract: **In this paper, we will explore the application of graph theory in mathematical biology. We will discuss the importance of the topic in the context of graph theory and will use graph theory to introduce basic biological networks. We will look at the application of graph theory in phylogenetic trees and further explore the use of trees in a more in-depth analysis of the gene sequence, BRCA1.**

Introduction

The choice to explore biological networks led to an attempt to find a bridge between graph theory and mathematical biology. On a broader scale, graph theory has many practical applications within networks in general. The concept came to me when we worked with trees in class. I used my knowledge in both graph theory and mathematical biology to investigate phylogenetic trees for relatedness between different species as well as prediction of mutations within gene sequencing.

What is a Tree?

For our purposes a tree, T, is defined as T= (V,E) where V is the set of vertices and E is the set of edges. The vertices in the tree are noted as nodes, and the edges are noted as branches. Beginning with the parent node, each branch below connects to an offspring. We will specifically be dealing with metric trees. In fig 1.1 we see a very basic example of a tree.

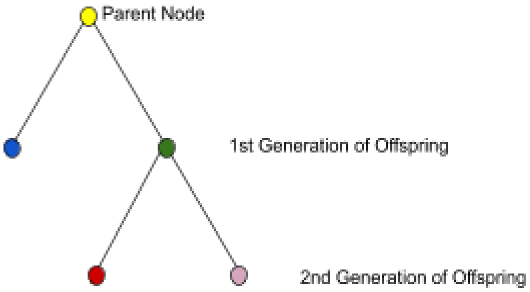


fig 1.1

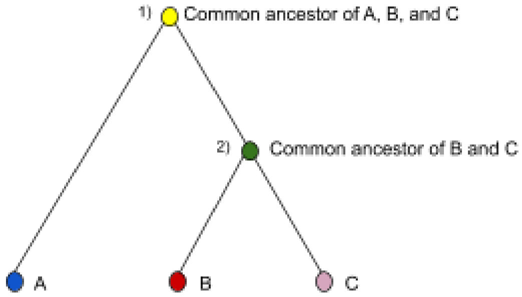

1) Common ancestor of A, B, and C
2) Common ancestor of B and C
A
B
C

Fig 1.2

Phylogenetic Trees

Phylogenetic trees are a visual representation of relatedness between different species.

On a basic level, phylogenetic trees are a visualization of biological diversity. See fig 1.2. Here we have modified fig 1.1 to fit a simplified version of a phylogenetic tree. The parent node (1) is the common ancestor that A, B, and C all share. (2) and A are the next generation of offspring of the parent node. B and C are the offspring of node (2).

From this tree, we are able to determine that B and C are more closely related than B to A, or C to A. Taking our phylogenetic tree one step further, we can create a line graph of fig 1.1, shown in fig 1.3. A line graph of any graph is the relation of edges to shared vertices. It is an easier visualization of the relations of A, B, and C to one another. We can then conclude from the line graph that the relation between B and C is closer than that of A. This concept can be extended to much more extensive relations between species.

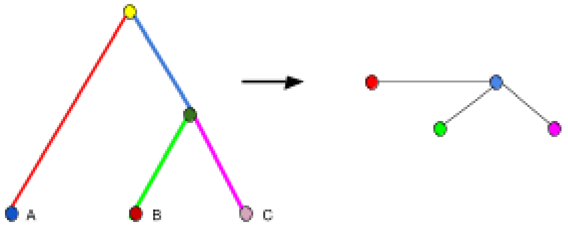


Fig 1.2

UPGMA (unweighted pair group method with arithmetic mean)

As an extension of our exploration of trees, UPGMA is a commonly used mathematical method used to create “biological trees” based on distancing and relatedness. There are many applications for this method. Aside from showing the relatedness between species by way of phylogenetic trees, because UPGMA is based in pair grouping, it can be used to show differencing by extension of arithmetic means. I decided to focus on the probability of mutation within gene sequences. This is also a slightly more interesting application of the process. It is a simple hierarchical clustering method, making for a relatively easy way to generate a “biological tree”. For this example, we will be looking at the BRCA1 gene.

Before we begin, a brief overview of the gene. BRCA1 is one of two well known “breast cancer susceptibility genes”, the other being BRCA2. BRCA1 is located on the 17th chromosome, and is known as a “tumor suppressor gene”. Mutation of either gene (BRCA1 or BRCA2) can cause higher susceptibility to breast cancer - as refers the name. When both genes function normally, they regulate the normal growth of cells within the breast. Mutations can allow for unregulated growth of cells, resulting in the potential formation of tumors.

Looking at an abbreviated gene sequence from BRCA1:

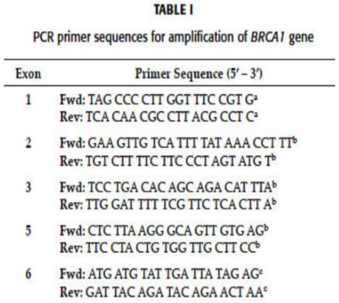
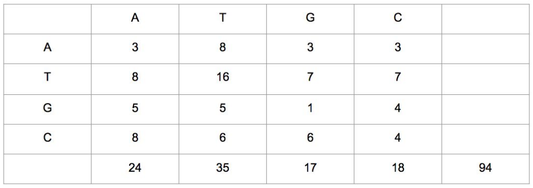


TABLE I

PCR primer sequences for amplification of *BRCA1* gene

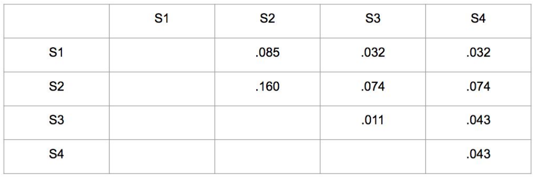
Exon	Primer Sequence (5′ – 3′)
1	Fwd: TAG CCC CTT GGT TTC CGT G[a] Rev: TCA CAA CGC CTT ACG CCT C[a]
2	Fwd: GAA GTTG TCA TTT TAT AAA CCT TT[b] Rev: TGT CTT TTC TTC CCT AGT ATG T[b]
3	Fwd: TCC TGA CAC AGC AGA CAT TTA[b] Rev: TTG GAT TTT TCG TTC TCA CTT A[b]
5	Fwd: CTC TTA AGG GCA GTT GTG AG[b] Rev: TTC CTA CTG TGG TTG CTT CC[b]
6	Fwd: ATG ATG TAT TGA TTA TAG AG[c] Rev: GAT TAC AGA TAC AGA ACT AA[c]

Fig 1.4



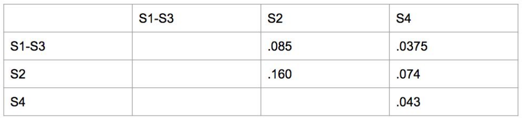
	A	T	G	C	
A	3	8	3	3	
T	8	16	7	7	
G	5	5	1	4	
C	8	6	6	4	
	24	35	17	18	94

Fig 1.5



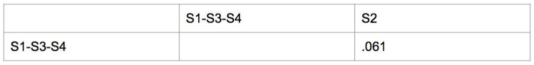
	S1	S2	S3	S4
S1		.085	.032	.032
S2		.160	.074	.074
S3			.011	.043
S4				.043

Fig 1.6



	S1-S3	S2	S4
S1-S3		.085	.0375
S2		.160	.074
S4			.043

Fig 1.7



	S1-S3-S4	S2
S1-S3-S4		.061

Fig 1.8

To find the probabilities:

1. add all individual terms in sequences aligned in chart divided by the total (number of “terms” multiplied by three) fig 1.5
2. take closest probabilities and combine two

terms into one term and found the distance of the new term in relation to the other probabilities, fig 1.6/1.7

3. repeat until all values are accounted for, fig 1.8
4. Use found probabilities to create tree, fig 1.9 and 1.10

The result is a metric tree representing the probability of mutation for each of the four terms in the gene sequence. This is a very relevant example of how we can use graph theory to help understand an otherwise unrelated use of a tree. In fig 1.10, I have reoriented the tree for easier understanding. The line on the bottom can be represented as "length". The shorter the distance, the lower the probability of mutation in that term. Thinking about the values as length can help to relate this tree back to fig 1.2, the basic phylogenetic tree.

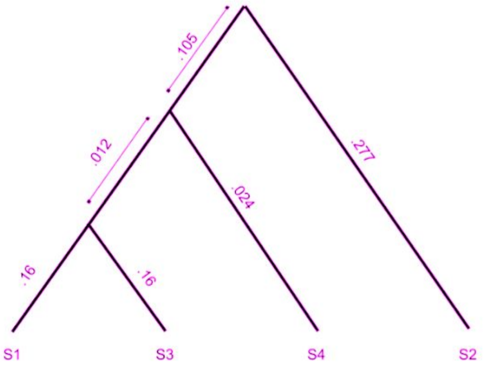


Fig 1.9

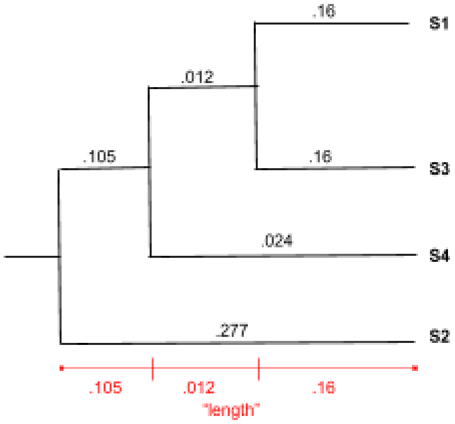


Fig 1.10

While gene sequence analysis is nothing new, it is pretty simple and cost-effective, thus it is an effective way to identify higher-risk patients. In a recent study conducted in 2009, 58 breast cancer patients were analyzed. Ten out of these 58 breast cancer patients carried *BRCA* mutations. Six (10.3%) in *BRCA1* and four (6.9%) in *BRCA2*. While this method is not perfect, I believe it is useful in countries that may not have the funding for extensive cancer prevention. Screening is relatively inexpensive, making for an accessible way to identify higher-risk patients. Additionally, with the rise in popularity of "at home" medical testing, I believe this method can be rendered into a product that would bring basic breast cancer susceptibility testing to the general public. I also believe it can be done at a relatively low cost, making it a good option for lower income households or those who do not have immediate access to medical care.

Conclusion

Graph theory, in general, has many applications. I believe this is true because most concepts in graph theory are relatively simple and can be used in conjunction with other mathematical processes to provide clarity and perhaps deeper understanding. By exploring phylogenetic trees, I was able to use the same concepts and apply them to UPGMA trees. This led me to an idea for a solution to a common problem, the problem being inaccessible healthcare. Overall, the use of graph theory in these specific instances allowed for clarity and further insight into the existing processes.

**Note: I originally wanted to explore biological networks, for example, population and predator/prey models. I believed there would be an obvious connection between the networks and that I would be able to represent these in a graph. However, I could not extend my research in social networking to relate enough to biological networks within my scope of knowledge in graph theory. I was able to explore within my original topic and find an application that was a better fit, which led to a more in depth look at what I ended up researching.

Additional sources:

https://jarhodesuaf.github.io/PhyloBook.pdf

https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0716-97602012000200003

https://www.maurerfoundation.org/the-breast-cancer-genes-brca1-brca2/