Build a tree to understand evolution

Bojana Đokić[1], Leon Konya[2] , Raffaele Masotti[3], Gracia Zrnić[4], Domagoj Gajski*

1 XIV. Beogradska gimnazija, Serbia
2 Bolyai Grammar School for Talents with Dormitory,Serbia
3 Calabrese Levi, Italy
4 Ženska opća gimnazija družbe sestara milosrdnica s pravom javnosti, Croatia
*Faculty of Science Zagreb, Croatia

Abstract

The aim of our project was to isolate DNA from different species of spiders and compare them to see how related they are. The first part consisted of collecting samples, isolating their DNA and amplifying parts of it. In order to do this, we used particular scientific methods from the field of molecular biology. In the second part we focused on building an evolutionary tree. We used an online database to find certain DNA sequences and phylogenetic software to align and compare them.

Keywords: alignment algorithms, evolutionary tree, gel electrophoresis, isolation of DNA, PCR, spiders

Introduction

People like to cut across natures complexity by sorting living beings into species. But how are all these creatures connected? Molecular phylogenetics is a branch of phylogeny which tries to gain information about organisms evolutionary relationships through their DNA. Phylogeny uses different classifications to build phylogenetic trees. The classifications come from morphological, biochemical, behavioural or molecular characteristics of species or other groups. In the past morphological trees were the most common way of classifying. Today we find that the most reliable technique is comparing the sequences of genes or proteins. Closely related species typically have few sequence differences, while less related species tend to have more.

Methods and materials

In the first part of our project we collected spiders. In order to do this, we went to a hill near Požega (latitude 45,3229404˚and longitude 17,67893˚), caught samples, put them in eppendorf tubes and after that we used them for extracting their DNA.

➢ **Digestion**

We put the spiders in Eppendorf tubes and added 200µL of Lysis Solution T which destroys the cell membrane and 20 µL of Proteinase K which disintegrated the proteins. Then we put the samples in water on 60˚C and left them there overnight.

➢ **Isolation and extraction:**

First we added 200 µL of Lysis solution C which inhibited the proteinase K because later on we will add polymerase (TAQ) which is a protein and proteinase K would disintegrate it as well.

Next step was preparing the column (filter) by putting 500 µL of CPS (Column preparation solution). That charged the filter positively and the negatively charged DNA will stick to it. Then we put our samples on the filters and centrifuged on 12 000 rpm. We added 200 µL of ethanol on the sample and centrifuged for 1 min on 10 000 rpm. And in the end we added 500 µL of washing solution on the sample and centrifuged for 1 min on 10 000 rpm. We repeated this step except we centrifuged it for 3 min. Then we centrifuged the sample by itself to get rid of any excess liquid. We eluted the DNA from the filter to the bottom of the tube by putting 50 µl of elution solution and centrifuging it on 10 000 rpm for 1 min (Figure 1.).

Figure 1: Our samples ready for centrifuge

➢ **Amplification**

Polymerase chain reaction (PCR) is a technique used in molecular biology for amplifying a single or a few copies of DNA segments.
In order to do this we made a mixture in tubes. Each student made one mixture:

- 0,75 μL of forward and reverse primers forH3 or S16 genetic markers. In two tubes we put H3 primers and in two tubes S16.
- 9 μL H2O
- 1,5 μL of DNA we eluted
- 15 μL PCR mix which contain polymerase (TAQ), dNTPs (deoxynucleotides), MgCl2, H2O and chemicals for detection of our samples on gel electrophoresis

After we finished, we put this solution into the PCR machine (Figure 2). The machine was programmed to have 3 phases of each cycle and a hot start. Hot start is necessary because of the activation of the polymerase. We used the TAQ polymerase (which is polymerase from thermophilic bacterium Thermus aquaticus) because this enzyme is able to withstand the protein-denaturing conditions caused by high temperature of the phases of PCR.

In the first phase the machine heats the solution to 95°C. At this temperature DNA denatures and we got 2 single stranded DNA molecules by breaking the hydrogen bonds between complementary bases.
In the second phase the machine cooled down to 50°C. At this temperature primers attached to DNA. Primers are short strands of RNA or DNA that serve as starting points for DNA synthesis.
The main difference between the forward and reverse primers is the direction in which they initiate the replication (Figure 3). The forward primer is complementary with the top strand (read from left to right) and the reverse primer is complementary with the lowest strand (read from right to left).

In the third phase the machine heated up to 72°C. At this temperature the Taq polymerase started to duplicate the DNA. Polymerase builds a new DNA strand by adding free dNTPs. The DNA copies are equal to the parental DNA.

Figure 2: The PCR cycler we used for our projects. Picture left represents the whole machine and the picture right are our samples prepared for the PCR process.

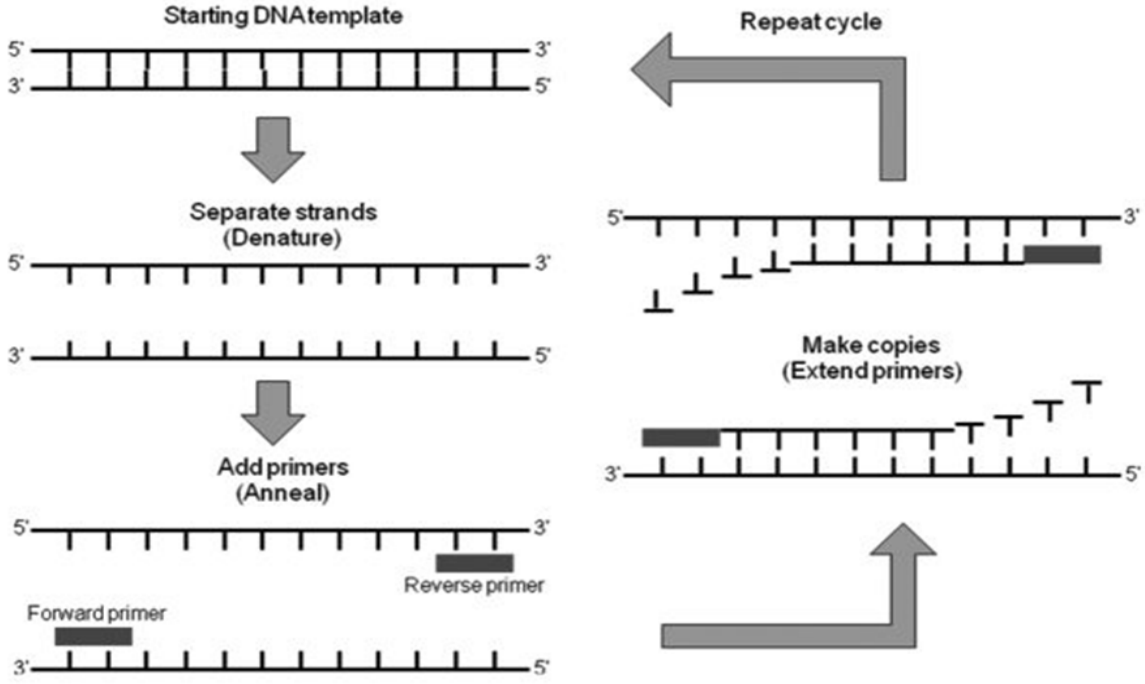


Figure 3: Polymerase chain reaction process

➢ Gel electrophoresis

Gel electrophoresis is a procedure we use to prove we have isolated double stranded DNA. First we made the gel by mixing 0.59 g of agarosis, 59 ml of TAE buffer (1% agarose gel) and adding a drop of sybrsafe. TAE buffer is a buffer solution containing a mixture of

Tris base, acetic acid and EDTA. Sybersafe is is a cyanine dye that attaches to double stranded DNA making it visible under UV light. It's important to be very careful with this chemical because it is cancerogenic. We microwaved, put it in a mould, removed bubbles and let it cool.

Then we put the eluted liquid from the PCR into indentations in agarosis gel and ran 60V of etlectric current through it. The negatively charged DNA travels through pores in the gel towards the positive electrode. After 15 minutes of running the current we placed the gel under UV light and saw dots appear at about the middle of the gel proving we isolated our pure fragments of DNA from genes H3 and S16without any other regions of the whole DNA we isolated (Figure 4).

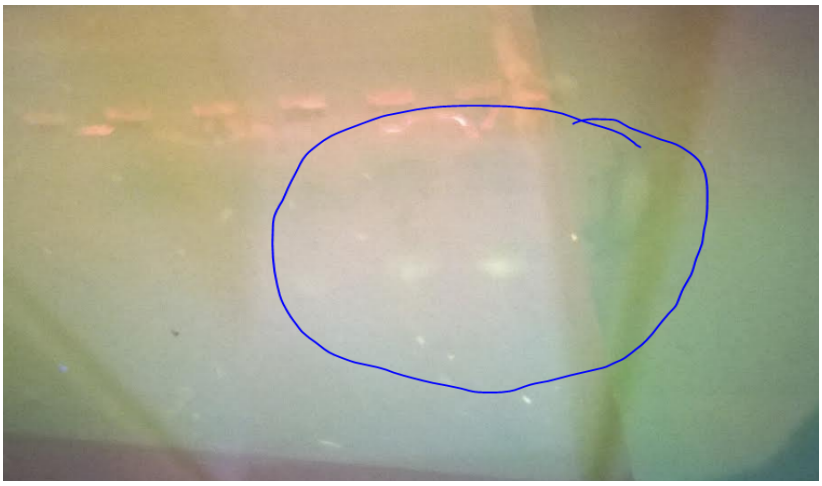


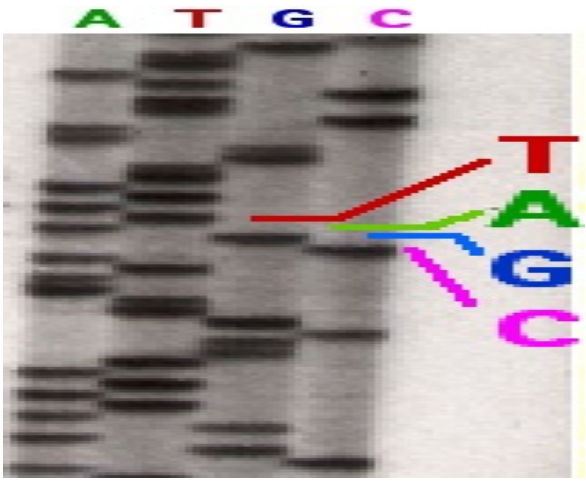
Figure 4: our amplified DNA fragments visible under UV light

- **Sanger sequencing**

If we had enough money and better equipment we would have done sequencing by using Sanger method. Sanger method was developed in 20th century and it was the most used method. Today we use something similar but the idea is the same.

For this method we need PCR mix and ddNTPs. ddNTPs (dideoxynucleotides) are like dNTPs (deoxynucleotides) except they will stop polymerase from building a second strand after their incorporation. The absence of the 3'-hydroxyl group means that, after being added by a DNA polymerase to a growing nucleotide chain, no further nucleotides can be added as no phosphodiester bond can be created.

We have four tubes with PCR mixture. In each we would add a small amount of different type of ddNTP, so we had tubes with ddATP, ddTTP, ddCTP and ddGTP, with the higher concentration of normal dNTP's. Now it's time for PCR. Because of ddNTPs polymerase will stop adding nucleotides after some time and we will have a lot of fragments with different length of DNA. When that's finished we would start Gel Electrophoresis. In each gel hole we would add mixture with different ddNTP. Electricity will go through gel and DNA will start moving and it will separate depending on size or length of the DNA. In the end we will have different positions of nucleotides and then we will just read them from bottom to top



(Figure5.).

Figure 5: DNA traveling trough gel electrophoresis for the Sanger sequencing method.We can see fragments of DNA separated by the size of only one nucleotide of difference. Picture was taken from the website: https://upload.wikimedia.org/wikipedia/commons/c/cb/Sequencing.jpg

➢ **Acquiring data**

Since the sequencing method was unavailable to us – lack of equipment and time, we couldn't use sequences based on our DNA samples. Because of that we used the most expansive genetic database, NCBI (National Center for Biotechnology Information) to download H3 and COI marker sequences of 20 species of spiders in FASTA format (Figure 6).

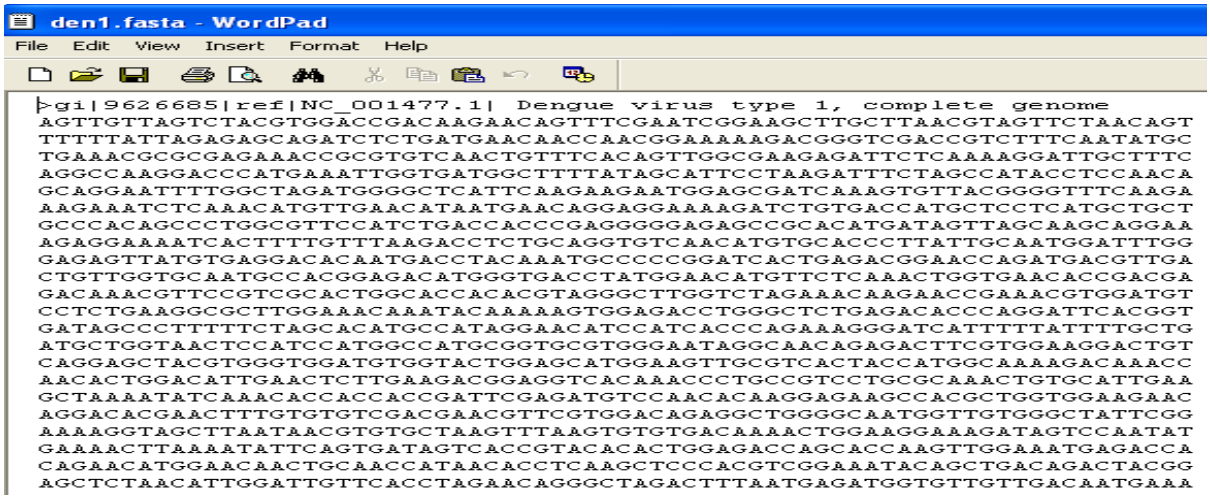


Figure 6: Fasta format sequence. Picture taken from the website: http://a-little-book-of-r-for-bioinformatics.readthedocs.io/en/latest/src/chapter1.html

In order to process the genetic information and to build the desired tree, we used the MEGA6 phylogenetic software (http://www.megasoftware.net/) , which is widely used by students and other people who want to learn more about phylogeny.

➢ Aligning DNA sequences

The next step was to align the DNA sequences so they would be more accurately comparable. We imported the FASTA formatted genetic information into MEGA6's alignment software. This software has different built-in aligners which use enhanced traditional, manual algorithms. The most accurate aligner is the ClustalW algorithm, which mainly relies on the Needleman-Wunsch algorithm. There's no point at learning how to align two DNA sequences manually since it's not efficient, but for the sake of insight, we learned this algorithm during the project.

The Needleman-Wunsch algorithm uses pre-made scoring matrices (there are some less accurate algorithms that calculate the temporary scoring table, but they require human intervention). That means that there is a certain value assigned to possible gaps (which are caused by deletions or insertions of nucleotides into our sequence of interest) in the sequence. Values have to be assigned to possible mismatches between nucleotides, and there are different values for different mismatch relations. This is because we differentiate two mutation mechanisms: transition and transversion (Figure 7). Transition happens between nucleotides with the same number of rings. Nucleotides from the purine group (adenine and guanine) are two-ringed, but from the pyrimidine group (cytosine and thymine) they are one-ringed. Transition happens more frequently because of the similarity between molecules with

the same number of rings – it is more likely to happen; therefore a transition mismatch in the substitution matrix has a lesser value than transversion.

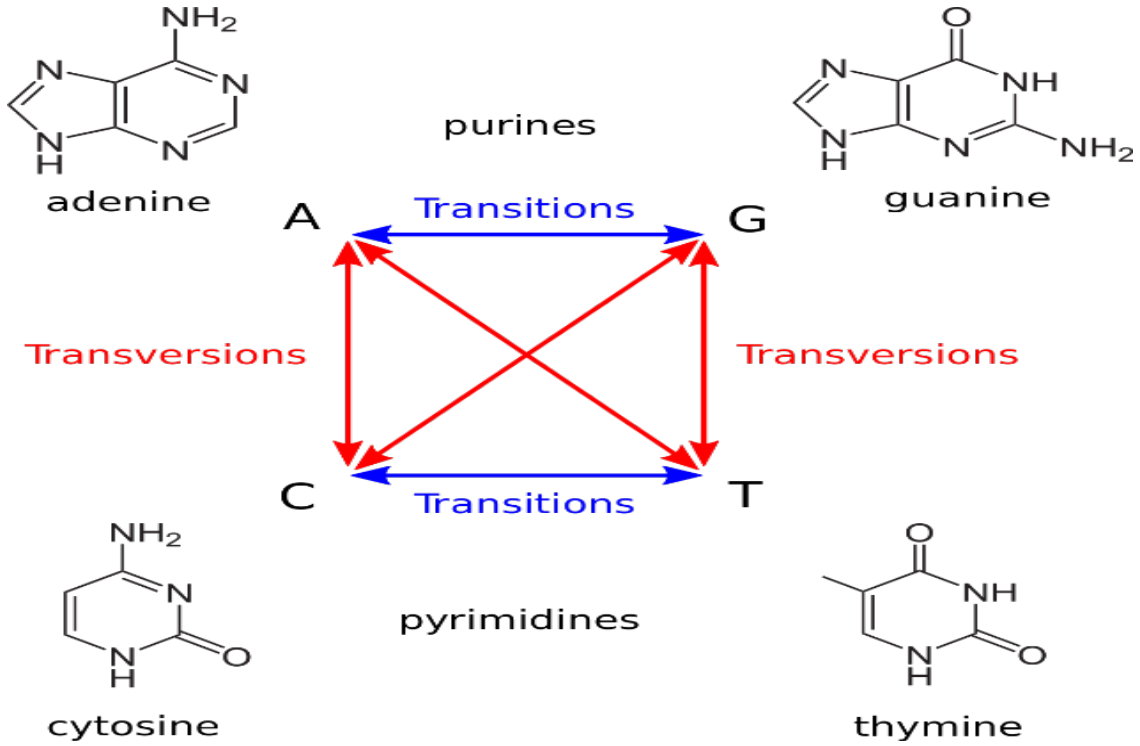


Figure 7: Types of mismatch mutations between nucleotides. Picture was taken from the website: https://en.wikipedia.org/wiki/Transversion

The Needleman-Wunsch algorithm is a dynamic one. It uses a matrix where the x and y angles represent the sequences that are being compared. The algorithm chooses the route with the least value from one end of the matrix to the other and reads the resulting sequence.

➢ **Building the tree**

Our goal was to find the most accurate model algorithm for constructing the phylogenic trees and understand the method. Because of this we used 3 different algorithms (UPGMA, Maximum parsimony and Maximum likelihood) and compared the results. In order to test the accuracy, we used outgroups – control sequences**:** H3 and CO1 genes of the fruit fly, which is relatively farther from the families of the chosen spiders.

➢ **UPGMA (Unweighted Pair Group Method with Arithmetic mean)**

This algorithm is the least accurate. It has been used decades before the first computers could compute it effectively, so it can be done manually. It is inaccurate because it uses a wrong hypothesis: that all mutations follow the same pace – this is called the molecular clock

hypothesis. Nonetheless, we learned this method because it was an important step in phylogeny and it was necessary for us to learn some terms and the jargon of this field.

The UPGMA method uses a distance matrix to create the tree. A distance matrix (Figure 8) shows the relative difference between each of the sequences, where the values of the differences = (number of different nucleotides)/(length of the sequences). The algorithm merges two sequences in the current matrix with the smallest difference and calculates each mean of the differences between the rest of the sequences and the two chosen sequences as the new distance value. This goes on until there is only one distance value left and the tree is assembled.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1. Araneus diadematus																					
2. Dolomedes tenebrosus	0.142																				
3. Loxosceles rufescens	0.167	0.179																			
4. Argiope lobata	0.034	0.152	0.178																		
5. Labulla thoracica	0.107	0.132	0.185	0.122																	
6. Ero sp.	0.117	0.122	0.133	0.136	0.099																
7. Cyclosa conica	0.047	0.153	0.157	0.074	0.108	0.103															
8. Filistata insidiatrix	0.151	0.171	0.170	0.167	0.159	0.134	0.154														
9. Cheiracanthium mildei	0.227	0.182	0.221	0.205	0.215	0.218	0.234	0.180													
10. Araneus marmoreus	0.012	0.148	0.172	0.047	0.111	0.121	0.051	0.150	0.239												
11. Zora spinimana	0.153	0.128	0.187	0.164	0.142	0.138	0.152	0.164	0.200	0.153											
12. Palpimanus sp.	0.184	0.193	0.241	0.178	0.224	0.224	0.210	0.216	0.261	0.180	0.202										
13. Textrix denticulata	0.184	0.143	0.188	0.166	0.156	0.154	0.172	0.188	0.214	0.173	0.165	0.211									
14. Clubiona huttoni	0.132	0.079	0.163	0.141	0.092	0.113	0.123	0.113	0.176	0.132	0.112	0.218	0.144								
15. Pisaura mirabilis	0.157	0.111	0.192	0.177	0.117	0.128	0.153	0.150	0.194	0.157	0.113	0.193	0.177	0.089							
16. Drapetisca socialis	0.174	0.132	0.159	0.183	0.074	0.133	0.142	0.190	0.200	0.168	0.178	0.242	0.166	0.122	0.168						
17. Floronia bucculenta	0.127	0.143	0.134	0.147	0.083	0.089	0.117	0.172	0.176	0.131	0.168	0.242	0.166	0.112	0.175	0.084					
18. Helophora insignis	0.190	0.164	0.132	0.187	0.139	0.153	0.158	0.179	0.227	0.194	0.148	0.236	0.148	0.122	0.140	0.143	0.153				
19. Argiope trifasciata	0.074	0.134	0.186	0.051	0.107	0.139	0.084	0.201	0.224	0.088	0.166	0.216	0.162	0.127	0.175	0.158	0.138	0.184			
20. Larinioides cornutus	0.052	0.134	0.173	0.088	0.108	0.128	0.047	0.166	0.260	0.056	0.164	0.205	0.208	0.122	0.164	0.159	0.138	0.191	0.089		
21. Drosophila melanogaster	0.228	0.224	0.214	0.216	0.205	0.224	0.222	0.230	0.256	0.211	0.234	0.255	0.252	0.198	0.209	0.213	0.244	0.226	0.235	0.233	

Figure 8: Distance matrix between our sequences of spiders for the gene H3. The 21st sequence ist from the fruit fly (*Drosophila melanogaster*) and represents our reference and outgroup.

➢ Maximum parsimony

The algorithm constructs a new tree at every nucleotide position, starting at the beginning and calculates the amount of mutations that have occurred according to the tree (Figure 9).

This algorithm ignores homoplasy – the case where the nucleotides changes back to its original base through multiple mutations – using the principle of the minimum evolution, which is similar to Occam's razor applied to phylogenetics. It states that having multiple hypotheses explaining something, the simplest explanation should always be the most viable. Using the principle, maximum parsimony algorithms select the constructed tree that has the least amount of mutations.

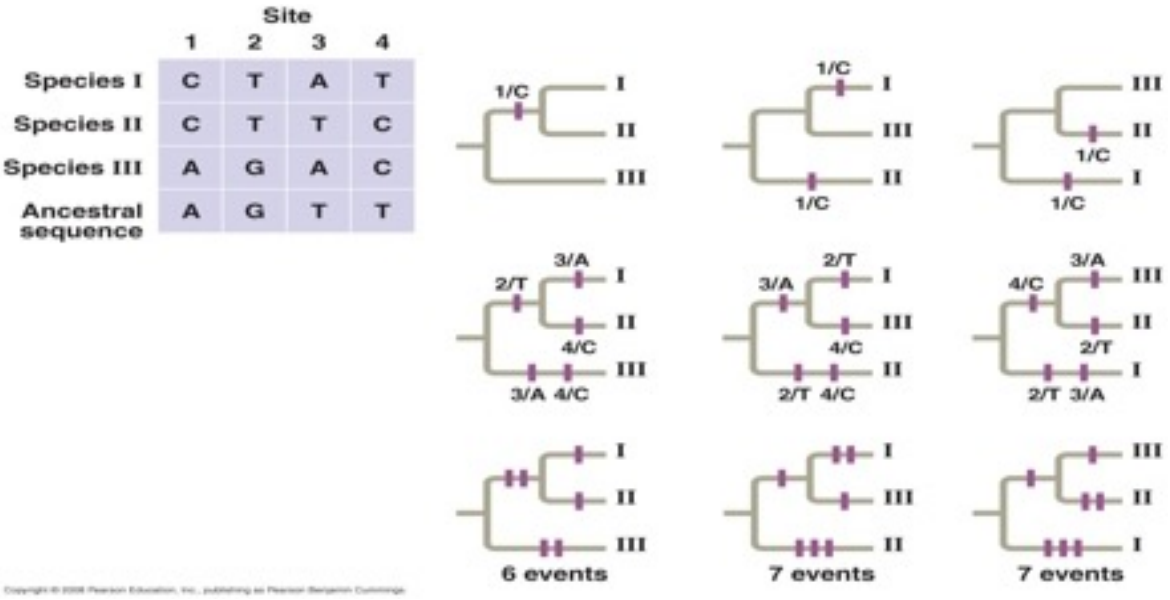


Figure 9: Example of the maximum parsimony method algorithm and how it builds and evolutionary tree. Picture taken from the website: http://www.assignmentpoint.com/science/psychology/maximum-parsimony-phylogenetics.html

Since the algorithm doesn't rely on any sophisticated mathematical formulas that are, to be efficient, only implementable in 4th generation computing, it can be considered obsolete and not very accurate. We made certain of this when we constructed the tree based on our collected data; the outgroup is racially not as distant from the spider genes as expected (picture 6.).

➢ Maximum likelihood

Maximum likelihood is the third method used to build trees and is most often used by today's researchers.

The maximum likelihood method uses standard statistical techniques for inferring probability distributions to assign probabilities to particular possible phylogenetic trees. The method requires a substitution model (in our case the K2 + G + I model) to assess the probability of particular mutations; roughly, a tree that requires more mutations at interior nodes to explain the observed phylogeny will be assessed as having a lower probability. This is broadly similar to the maximum-parsimony method, but maximum likelihood allows additional statistical flexibility by permitting varying rates of evolution across both lineages and sites. In fact, the method requires that evolution at different sites and along different lineages must be statistically independent.
The more probable the sequence given the tree, the more the tree is preferred. All possible trees are considered and because of that it is computationally intense and takes a lot of time to calculate the most probable tree. For example, it took us with the computers at the school

only one minute to calculate the UPGMA and maximum parsimony tree, but for the maximum likelihood tree, with the same amount of data, it took the computer 2 hours to calculate a tree.

When using a exhaustive method like maximum likelihood it is always good to simultaneously run the bootstrap calculating algorithm. In the statistical context, bootstrapping refers to using the data at hand to infer the uncertainty of said data. I.e. improve the statistic by pulling on its bootstraps. In practice, this is achieved by sampling or permuting the input data. In terms of a phylogenetic tree, the bootstrapping values indicate how many times out of 100 the same branch was observed when repeating the phylogenetic reconstruction on a re-sampled set of your data. If you get 100 out of 100 (and your data is sufficiently large to support this), we are pretty sure that the observed branch is not due to a single extreme datapoint. If you get 50 out of 100, we cannot be as certain.

Results and discussion

- **UPGMA**

We built our tree with this method using the Mega6 phylogenic software. Even though the outgroup was the least related to the spider sequences, the results (Figure 10) still proved that the algorithm is inaccurate. For instance, at least two spiders from the same family (the family relatedness was concluded through morphological and phylogeny research before) have been listed as not belonging to the same family (Argiope family). This method is only useful for closely related groups of organisms whose sequences haven't changed dramatically because the distance matrix is affected by homoplasy and encounters errors if the percentage of similarity between sequences is too low.

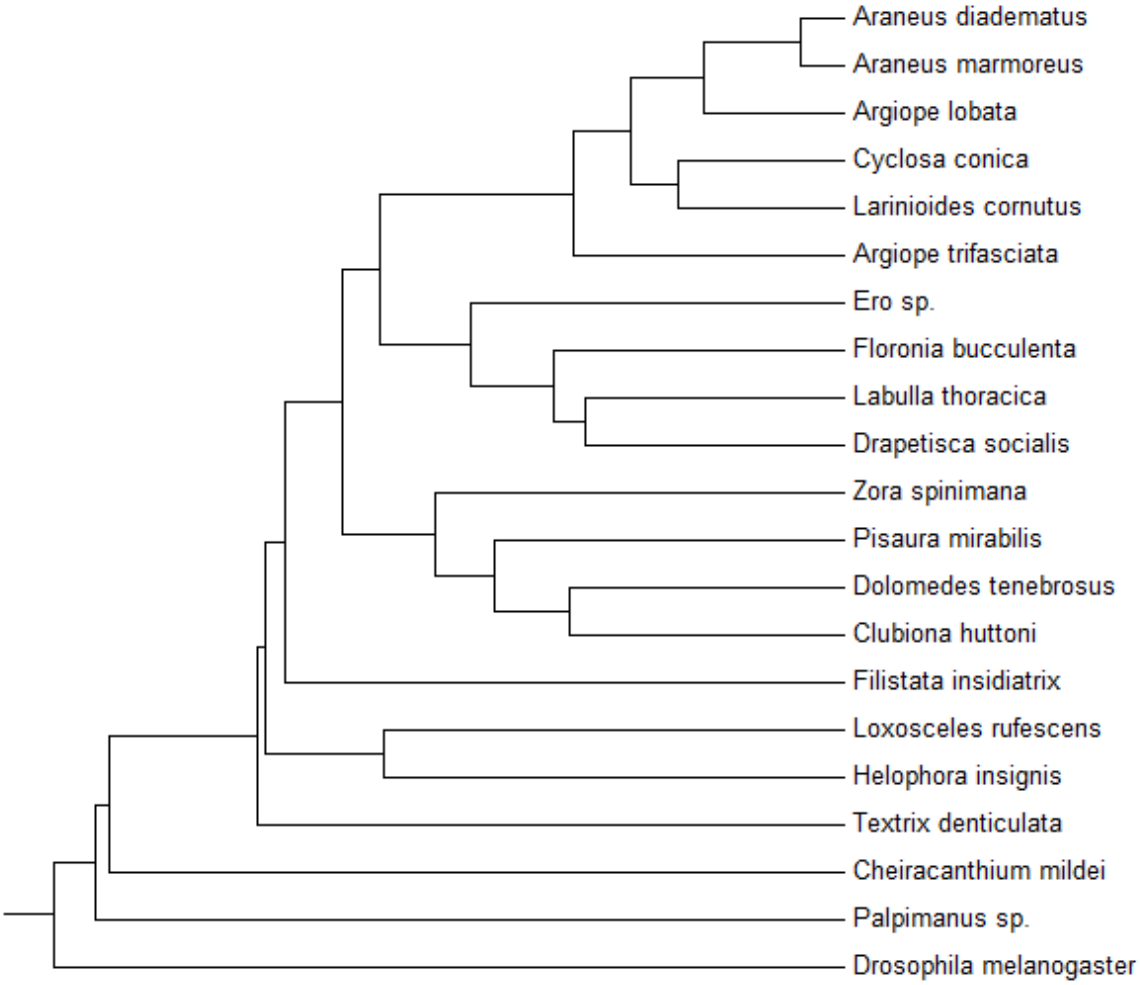


Figure 10: Phylogenetic tree based on our collected data from spiders build by the method UPGMA (Unweighted Pair Group Method with Arithmetic mean).

➢ **Maximum parsimony**

Since the algorithm doesn't rely on any sophisticated mathematical formulas, it can be considered obsolete and not very accurate. We made certain of this when we constructed the tree based on our collected data; the outgroup (*Drosophila melanogaster*) is racially not as distant from the spider genes as expected (Figure 11).

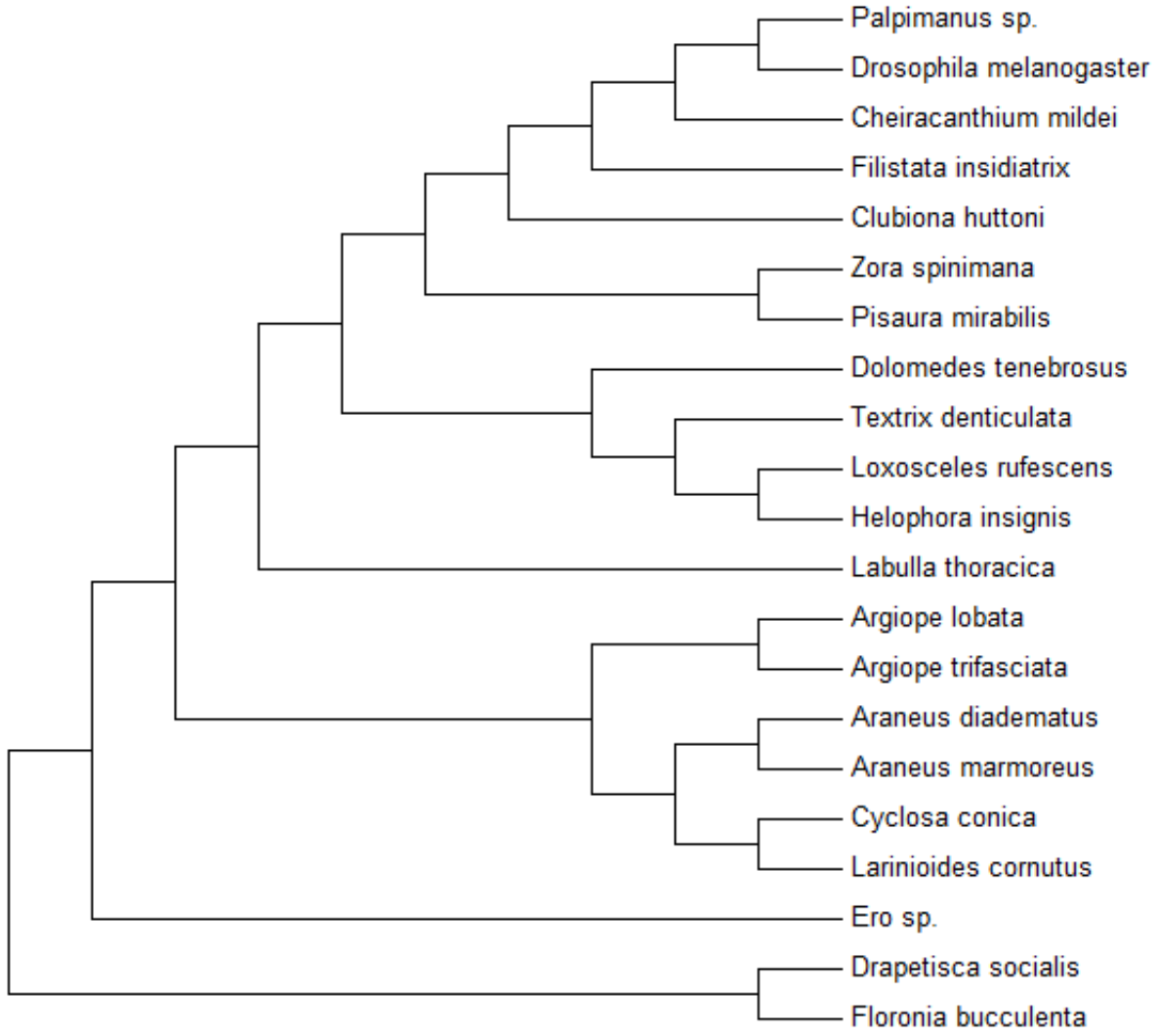


Figure 11: Phylogenetic tree based on our collected data from spiders build by the method maximum parsimony.

➢ **Maximum likelihood**

As we can see from Figure 12, even with the most exhaustive and precise method for building evolutionary trees, we can sometime get confusing results. Even here are the closely related species from the same family or order separated by more than one node. The bootstrap values are quite small for almost all of our nodes and branches.

With such results, we came to the conclusion that we lack a lot of data and need all the species form all the families to talk about statistically significant results. Not only do we need all the species, but we need data of even more evolutionary markers of genes which we could compare so that the tree gets a higher bootstrap value of confidence.

We also have to take the problem of homoplasy into account, meaning that the rate of substitutions calculated by the model algorithm can give us the worst model for our set of

data. It's because the nucleotides can mutate again to the same nucleotide and we can't detect such changes with our current methods.

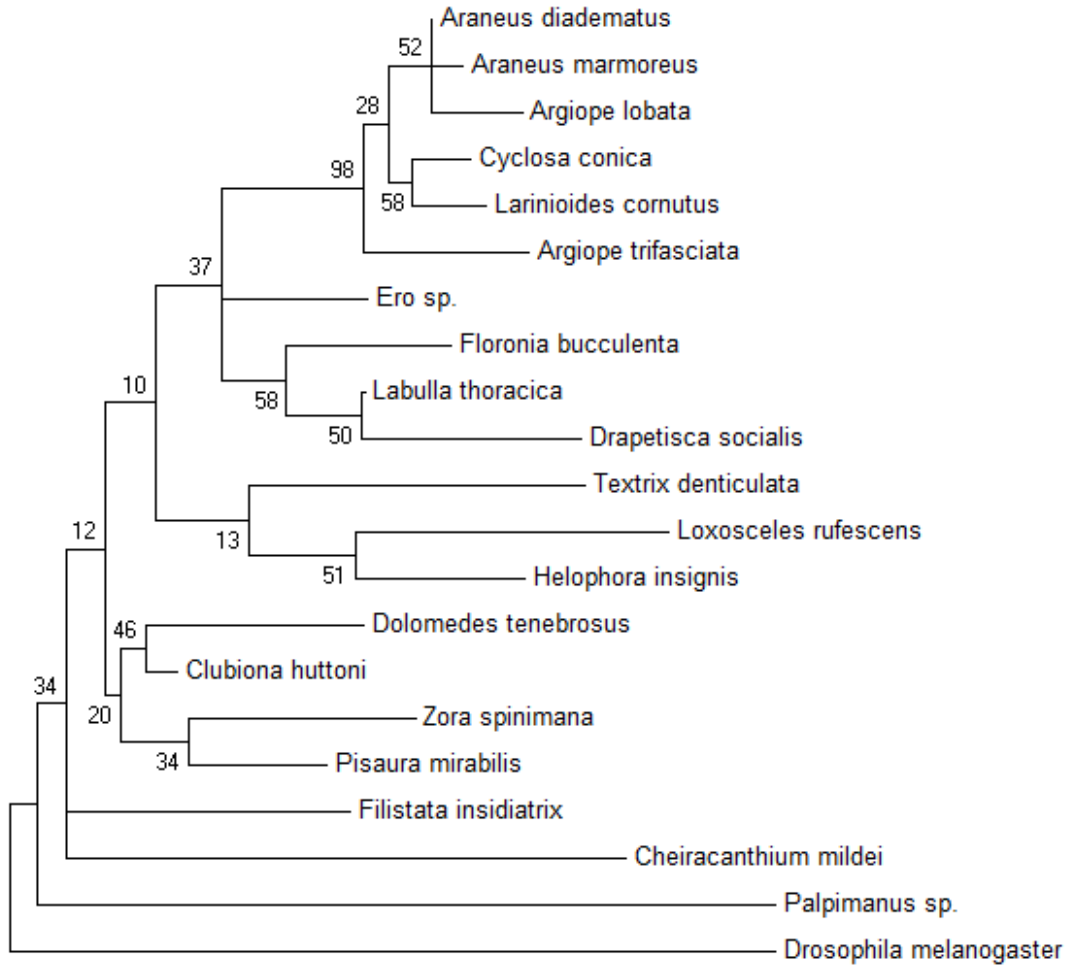


Figure 12: Phylogenetic tree based on our collected data from spiders build by the method maximum likelihood with the bootstrap values of each node.

Conclusion

Even though we used 3 different algorithms, we can't precisely tell which species are closer to other ones. Each algorithm has lapse or leave out specific factor of vital significance for comparison. Moreover to build valid evolutionary tree we need to have much more samples.

Acknowledgments

This project was supported by doc. dr. sc. Ivana Ivančić Baće, and doc. dr. sc. Branka Bruvo Mađarić from the Faculty of science in Zagreb by giving us all the equipment we needed.

We thank Dora Grbavac (technical support), Nikolina Šoštarić (organiser) and Sebastijan Dumančić (organiser) for assistance with obtaining all the devices we needed to fulfil our tasks for the project.

References:

1. https://en.wikipedia.org/wiki/Computational_phylogenetics
2. https://www.ncbi.nlm.nih.gov/