Instituto Politécnico de Setúbal

Escola Superior de Tecnologia do Barreiro

Biological Sequence Analysis

2021/2022 Bioinformatics
Prof. Fransisco Pina Martins

Assignment 1
"Knee deep into phylogenetics"

May 2022

Group:
Carolina Rodrigues, nr 201900225
Guilherme Sá, nr 202000201
Nuno David Sardinha, nr 201900192
Francisco Amaral, nr 201900202

Table of contents

List of Figures

1 Introduction

This group work intends to make us go "knee deep into phylogenetics", meaning, to give us a better understanding of the phylogenetic analysis procedure, as well as to improve our knowledge on the content taught in the practical lessons of the biological sequence analysis unit.

Phylogenetic analysis refers to studying or investigating the evolution of a species or group of organisms or a feature of an organism. Phylogenetics are important because it helps us understand how genes, genomes and species evolve.[1]

For this work, first, we must look for an article that carries out a phylogenetic analysis. Once we find one that matches the requirements, we have to understand their proposed biological problem, then, obtain the sequences used, in order to reproduce the analyses that were carried out in the paper by using the methods that we find most relevant/important, and finally, we must interpret the phylogenetic trees produced and compare the results with those in the original article.

The chosen article is "Phylogeny and differentiation of the St genome in *Elymus* L. sensu lato (Triticeae; Poaceae) based on one nuclear DNA and two chloroplast genes".[2]

Figure 1: Example of one of the *Elymus* L. species used in this study. (font: *Elymus sibiricus*)

1.1 The original biological problem:

Upon reading this paper, we understood that the purpose of the research group was to analyze the mechanisms of hybridization and polyploidization in the evolution and specification of plants, to elucidate phylogenetic relationships in certain *Elymus* polyploids, investigate the genetic differentiation of the St genome in *Pseudoroegneria* and to study the genetic differentiation of the St genome in the polyploid *Elymus* sL.

"Polyploidization", or polyploidy, is the presence of species with a common multiple chromosomal number. It is closely linked to several key areas of research, including plant species diversification, innovation in genetic function, and the development of important traits.
Polyploidization is seen as the main driving force behind the diversification of plant species and plays an important role in the evolution of the plant genome and its genetic improvement. It results in multi-copy genes that exist as paralogs to each other. [3]

The term "hybridization" suggests the genetic crossing between two species and has the fundamental role of shaping the history of life on earth and in the development of several lineages. It is deliberately used in the selection of domesticated plants to take advantage of transient hybrid vigor, displace desirable variation among lineages, and generate new phenotypes.
Hybridization is often considered only between species, but "from a genetic point of view, interspecific hybridization is only a special case of a much more widespread phenomenon" (Stebbins, 1950).
This is evident, for example, in interspecific hybridization of wheat and cotton species to form allopolyploids (allopolyploidy is an evolutionary process through which two or more different genomes are united into the same nucleus).[5]

Elymus L. sensu lato species are allopolyploids that assumed to share a common St genome of *Pseudoroegneria*, in different combinations with the H, Y, P and W genomes. However, as the genome of St has evolved in this species, during the processes of hybridization and polyploidization, it is not yet clear whether this genome matches the pattern for this species.

So, since hybridization and polyploidization may be the main mechanisms for plant evolution and speciation, the polyploidization process and the history behind the evolution of polyploids is of great interest for a better understanding of the evolutionary history of the Plantae kingdom.

1.2 The original methods used:

The methods used in the original paper, were: taxon sampling, DNA extraction, amplification and sequencing, phylogenetic analysis, network analysis and Nucleotide diversity estimate. (All the following abbreviations and therms used

in this assignment, regarding the methods used in the original paper, can be accessed here - [2])
For the analysis, the authors used 6 accessions of 4 *Pseudoroegneria* species with St genome, 35 accessions of 12 other diploid species with monogenome P, W, V, H, I, E, XP, NS and 28 allotetraploid *Elymus* sL, using a transcribed spacer region of the nuclear gene and two chloroplast genes. They obtained the maximum likelihood (ML) tree of the nuclear ribosome and chloroplast ML tree.

- **Taxon sampling**

Taxon sampling refers to the process of selecting representative taxa for a phylogenetic analysis.
For this study, they included 28 *Elymus* s. l. species and were analyzed together with 16 diploid taxa. *Bromus inermis Leyss* was used as outgroup. And then, were collected the seed materials with Pr, ZY, and Y numbers.

- **DNA extraction, amplification and sequencing**

DNA extraction is a method, by using physical and/or chemical methods from a sample to separate DNA from cell membranes, proteins, and other cellular components. The amplification refers to the increase in the number of copies of a segment of DNA. And, the DNA sequencing, refers to the technique used for determining the exact sequence of nucleotides, or bases, in a DNA molecule.

In the study present on the original paper, the CTAB (Cetyltrimethyl Ammonium Bromide) procedure was used to isolate total DNA. The nuclear nrITS sequence, chloroplast matK and trnH-psbA spacer sequence were amplified with the primers ITS4, ITS5, W, 9R, trnH1 and trnH2.

PCR amplification of the cpDNA was carried out in a 50 L reaction mixture, containing 10× ExTaq polymerase buffer, 2 mM MgCl2, 200 M of dNTP, 1 M of each primer, 1.5 U ExTaq and about 30 ng of template DNA.

Amplifications were performed on Mastercycler (Pro S,Eppendorf, Germany) using the following gene Protocols:
for nrITS - 1 cycle: 5 min 95 °C; 35 cycles: 1 min 94 °C, 1 min52 °C, 1 min 72 °C; 1 cycle: 8 min72 °C.
For matK - 1 cycle: 4 min 95 °C; 35 cycles: 1 min 94 °C, 1 min50 °C, 1.5 min 72 °C; 1 cycle: 10 min72 °C.
And **for trnH-psbA** - 1 cycle: 4 min 95 °C; 25 cycles: 1 min 94 °C, 1 min56 °C, 1 min 72 °C; 1 cycle: 7 min72 °C.

The PCR products were visualized on 1.0 % agarose gels, purified by an ENZA™ gel extraction kit and then cloned into pMD19-T vector. Three random clones per diploid were chosen to sequence. As there are at least three to five accessions for each allopolyploid in this study, only one random clone for each accession of allopolyploid was picked and sequenced.

- **Phylogenetic analysis**

The multiple sequences alignments were made using ClustalX and the phylogenetic analyses were performed using Maximum likelihood (ML) method. Maximum likelihood analyses of the nrITS data, matK data and trnH-psbA data were performed in PAUP*4.0b10.

To determine the best evolutionary model to be used for the phylogenetic analyses, they used the ModelTest v3.0 with Akaike information criterion (AIC). The optimal model were GTR + G for nrITS data, TVM + G for matK data, and K81uf + G for trnH-psbA data.

Maximum likelihood heuristic searches were performed with 100 random addition sequence replications and Tree Bisection-Reconnection (TBR) branch swapping algorithm. In order to infer the robustness of clades, bootstrap support (BS) values were calculated with 1000 replications.

- **Network analysis**

The network analysis (NA) is a set of integrated techniques to depict relations among actors and to analyze the social structures that emerge from the recurrence of these relations. Taking into consideration the potential for reticulation in the evolution of polyploids, phylogenetic network reconstruction method was used to study the relationship between ancestral and derived haplotypes in this study.

For this simulation, they used known gene genealogies and the median-joining (MJ) network method was performed. The MJ network analysis was generated by the Network 4.6.1.3 program. Because the program infers median-joining networks from non-recombining DNA, the GARD recombination detection method within the HyPhy package was used to test for recombination.

- **Nucleotide diversity estimate**

Nucleotide diversity estimate is a measure to examine the genetic variation. For this study, to assess the gene divergence and genetic relationships in the St genome, between polyploids and its diploid progenitor, nucleotide diversity was estimated by using Tajima's (pi), and Watterson's (theta).

Tajima's (pi) quantifies the mean percentage of nucleotide differences among all pairwise comparisons for a set of sequences, while Watterson's (theta) is simply an index of the number of segregating (polymorphic) sites.

Tests of neutrality including Tajima's and Fu and Li's D statistic were performed. Significance of D-values was estimated with the simulated distribution

of random samples (1000 steps) using a coalescence algorithm assuming neutrality and population equilibrium. These parameters were calculated with DnaSP 4.10.9.

1.3 The conclusions from the original paper:

Based on the obtained results, the authors concluded that there is a clear connection between the polyploid species sequences of *Elymus* sL and those of its diploid ancestors.

The combination of this, with previous cytogenetic results, supports the hypothesis that *Pseudoroegneria*, *Hordeum* and *Agropyron* species served as diploid donors of the St, H and P genome during the polyploid mechanism of *Elymus* sL species specification. Which also proves that the evolutionary differentiation of the St genome in *Elymus* sL, followed by the emergence of this group, may be caused by mechanisms of hybridization and polyploidization.

Those results suggested that the ancient common maternal ancestral genome in *Elymus* s. l. is the St genome from *Pseudoroegneria*.
The ancient genome of the common maternal ancestor of *Elymus* sL is the genome of St de *Psedoroegneria*.
Based on the ML tree analysis, they speculated that *Elymus* sL species originated from Central Asia and Europe, spreading to North America.

1.4 The goals of our work:

The main purpose of this work, after the reading and understanding of the original article, is to reproduce the phylogenetic analysis performed (except for the Median-joining (MJ) network methods used in the Network analysis), making sure those are reproducible.
After, we should interpret and compare our results to those in the original article.

2 Methods and Tools

To perform this work, and for all the procedures, a virtual machine (VM) was used, which is a virtual environment that simulates a physical machine. We used the Oracle VM VirtualBox 6.1.34 version to run the VM with an optical drive Ubuntu 21.10 version.

Ubuntu [5] is an open-source and freely available operating system, one of the Linux [6] distributions.
VirtualBox is designed to execute virtual machines on a physical machine without reinstalling the operating system (OS). Oracle VM VirtualBox is cross-platform virtualization software that enables users to expand their existing computer to run multiple OS at the same time. An OS is a software which manages all the hardware resources associated with the desktop or laptop.[7]

(a) VirtualBox logo

(b) Ubuntu logo

(c) Linux logo

For the management and storage of files and scripts needed in this work, we used GitHub.
GitHub is a cloud storage code hosting platform that allows programmers to collaborate and make changes to shared projects from anywhere. [8]
For the purpose of making accessible all the content required for this assignment, we created a repository, available at GitHub - Assignment-ASB [9]

After choosing and reading of the original article "Phylogeny and differentiation of the St genome in *Elymus* L. sensu lato (Triticeae; Poaceae) based on one nuclear DNA and two chloroplast genes" [2], was time to reproduce the analysis therein performed.

• 2.1 Sequences collection and alignment

To initiate the phylogenetic analysis, we had to obtain the sequences used in the original article.
Consulting the "Additional file 1: Table S1" [9] we managed to download the XML file, containing the "Table S1: Species of *Elymus* sensu lato and the related species used in this study", that would be our data set to be used during our analysis.

To collect the sequences, we created a script using shell programming language [10], “lemanel2.0.sh” (available in the GitHub - Assignment-ASB repository), It basically, intends to search and collect from National Center for Biotechnology Information (NCBI) [11] the data of all the genomes and sequences needed for the replication of the article, aligning those sequences and storing them in a FASTA format file.

To execute the script, using the command line of the terminal, we must put the following instruction: “./lemanel2.0 database query” (as exemplified in the [Figure3]).

’database’ represents the first argument, in this case was used “nucleotide”, the second argument and ’query’ representing the Accession no. of the species (values available in the original paper “Additional file 1: Table S1” previously downloaded).

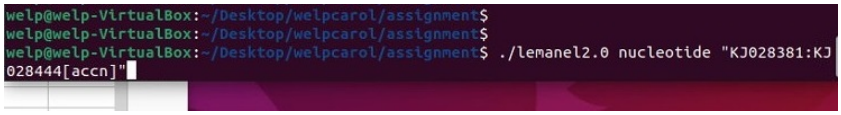

Figure 3: Practical example of how to execute the lemanel2.0.sh script in the terminal.

The code “lemanel2.0.sh” consists of 3 main functions: Esearch(), Efetch() and Remove().
The **Esearch()** function, after writing the intended database and the query argument, gets from the NCBI the respective ”query_key” and the ”webenv”, to be used in thw Efetch function.
The **Efetch()** function, that uses them to, again from the NCBI, get the desired sequences, and store them in a FASTA file. Using MAFFT (Multiple Alignment using Fast Fourier Transform) [12] for the alignment of the previous sequences, and creating a new file to store them.
And finally, since that in the original paper the trees have only the species name and the sequence number, we have to filter the names of each sequence. That’s where The **Remove()** function comes. It runs the lines of the aligned sequences file, and filters by gene the names of each sequence, in order to leave only the relevant part of the name (specie and sequence number).

(a) GitHub logo (b) NCBI logo (c) Bash logo

- Problem encountered / Solution found

We intended to extract only 3 times the sequences, one for each gene (ITS, matK, trnH-psbA), changing only the second argument, in order to generate the 3 files needed to be used in further steps.
But, and since we couldn't manage to find a way to write the 2nd argument all at once (due to empty cells in the original table) we had to, for each gene, extract 2 files. One, we can call it "fasta X" that represents the sequences obtained from the data that was sequential, and the other, we can call it "fasta Y" represents the alternated data.
For each pair of fasta X and X files, we used MEGA to join them and align with MUSCLE (our option), to create a single sequenced FASTA format file for each gene.
MEGA (Molecular Evolutionary Genetics Analysis) is a software program available to help both scientists and students in making dendrograms or phylogenetic trees, using nucleotide or protein sequences. And for this work we used the MEGA X version 10.2.4. [13]

• 2.2 Tree implementation

To create the trees, first, we should test the best model to build the tree. For that, in MEGA, after importing the completed aligned files, used the option "Find Best DNA/Protein Models (ML)...". As shown in [Figure 5].

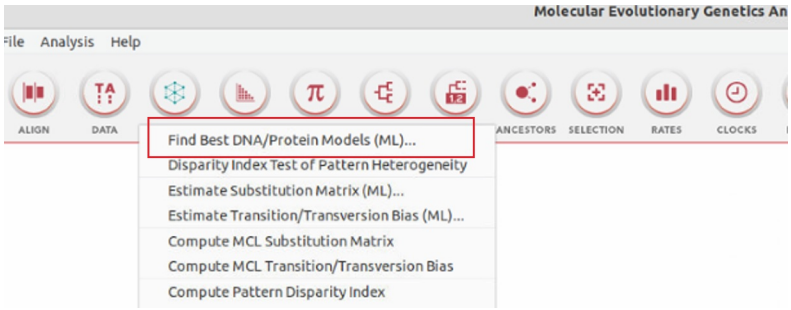

Figure 5: Illustration of the "Find Best DNA/Protein Models (ML)..." option in MEGA software.

This option generated a table [Figure 6], from which we can observe and chose the model we find more adequate to use for the building of the trees.

To build the trees, we used Ugene software, version 42.0.
Ugene is a multi-platform available for genome analysis with various tools for the desired purposes. It supports various data formats, visualization and editing of DNA and protein sequences, among others. [14] We must use the terminal to open this software, as illustrated in the 7].

Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	*lnL*	*(+I)*	*(+G)*	*R*	*f*(A)	*f*(T)	*f*
T92	187	8200.617	6540.105	-3082.392	n/a	n/a	1.20	0.322	0.322	0.
T92+G	188	8210.304	6540.919	-3081.793	n/a	7.71	1.24	0.322	0.322	0.
T92+I	188	8210.322	6540.938	-3081.802	0.00	n/a	1.20	0.322	0.322	0.
T92+G+I	189	8219.506	6541.249	-3080.950	0.00	6.69	1.24	0.322	0.322	0.
HKY	189	8222.411	6544.153	-3082.403	n/a	n/a	1.21	0.347	0.298	0.
TN93+G	191	8228.920	6532.918	-3074.771	n/a	8.61	1.24	0.347	0.298	0.
HKY+I	190	8231.924	6544.794	-3081.716	0.00	n/a	1.21	0.347	0.298	0.
HKY+G	190	8233.236	6546.106	-3082.372	n/a	133.79	1.21	0.347	0.298	0.
TN93	190	8233.242	6546.112	-3082.375	n/a	n/a	1.21	0.347	0.298	0.
HKY+G+I	191	8241.164	6545.162	-3080.892	0.00	6.83	1.24	0.347	0.298	0.
TN93+I	191	8242.794	6546.791	-3081.707	0.00	n/a	1.21	0.347	0.298	0.
GTR	193	8250.717	6536.970	-3074.782	n/a	n/a	1.29	0.347	0.298	0.
TN93+G+I	192	8252.036	6547.161	-3080.885	0.00	6.83	1.24	0.347	0.298	0.

Figure 6: Table generated with MEGA, after "Find Best DNA/Protein Models (ML)..." option was applied.

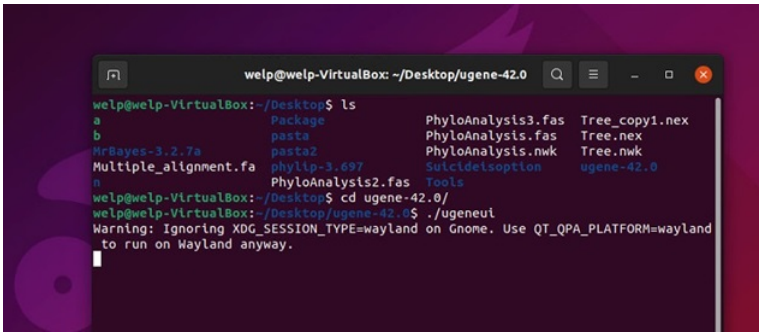

Figure 7: Instructions needed to put in the terminal, in order to open Ugene software.

Once the software is open, and the final aligned sequences imported, we noticed that the Neighbor joining option was chosen automatically, so, we change it to use Maximum likelihood instead. Then, we select the best model, as previously established.

• 2.3 Production of the report

For the writing of the report, was used Overleaf, which is an online writing tool to create LaTeX documents in the web browser, and can be shared in order to have multiple collaborators, as we did.
LaTeX tool is made to create documents in a much faster and more intuitive way, by writing and editing code in a window, and the automatic formated document can be visualized in the other, the word "LaTeX" is an abbreviation of "Lamport's TeX", named after Leslie Lamport. [15]

3 Results and discussion

The raw trees produced by Ugene software, were monochromatic, predefined with black color. So, to better compare them with the trees used in the original article, we had to color them.
This was, in our opinion, one of the most tedious part, because we didn't find a way to automate it, so it had to be done to all the species, one by one, while checking, simultaneously, the original paper, so that we could match the colour of the species.

- nrITS analysis

In the original article, the likelihood settings from best-fit model was GTR + G. (Selected by Akaike information criterion (AIC) in Modeltest 3.7.)

In our replication of the nrITS analysis [Figure 8], even if the best model in the original paper was GTR+G, due to limitations in the software Ugene, was not possible to select it. Thus, we ended up using only the GTR, a more generic model. Because of that, the configuration of the tree ended up being slightly different. The clades are still visible, however, located in different positions.

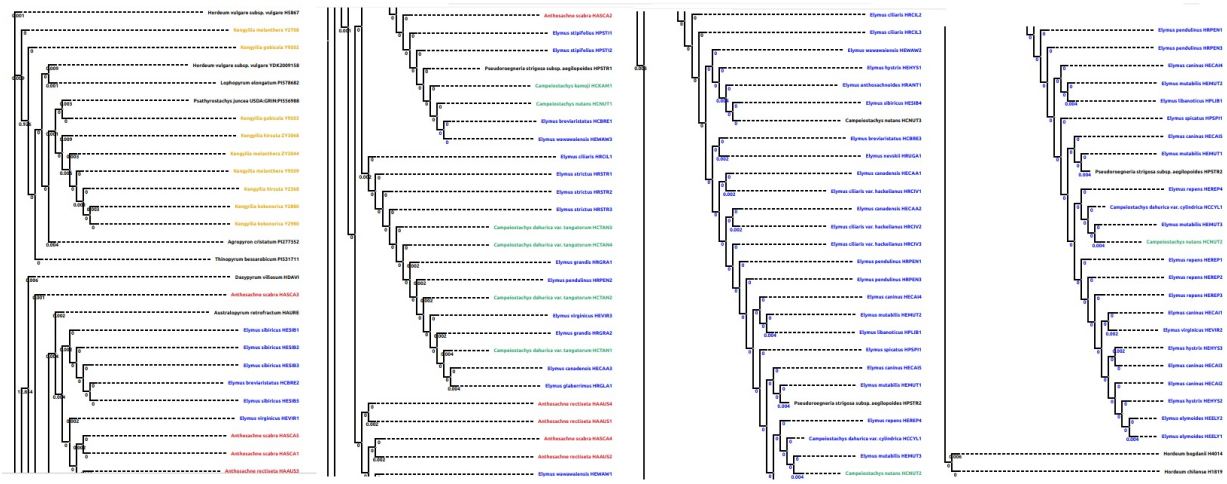

Figure 8: Our reproduction of the ntITS tree, using Ugene software.

- matK analysis

In the original article, the ML analysis of the matK sequence data yielded a single phylogenetic tree. Likelihood settings from best-fit model was TVM + G, selected by AIC in Modeltest 3.7. In their tree for the matK, all the *Elymus* s. l. species and some diploid species of the *Triticeae* formed Clade I.

In our replication the matK analysis [Figure 9] and [Figure 10], even if the best model in the original paper was TVM + G, due to limitations in the software Ugene, was not possible to select it. Thus, we ended up using the TVM, a more generic model. Because of that, the configuration of the tree ended up being different, as the clades are still visible, however, in different positions.

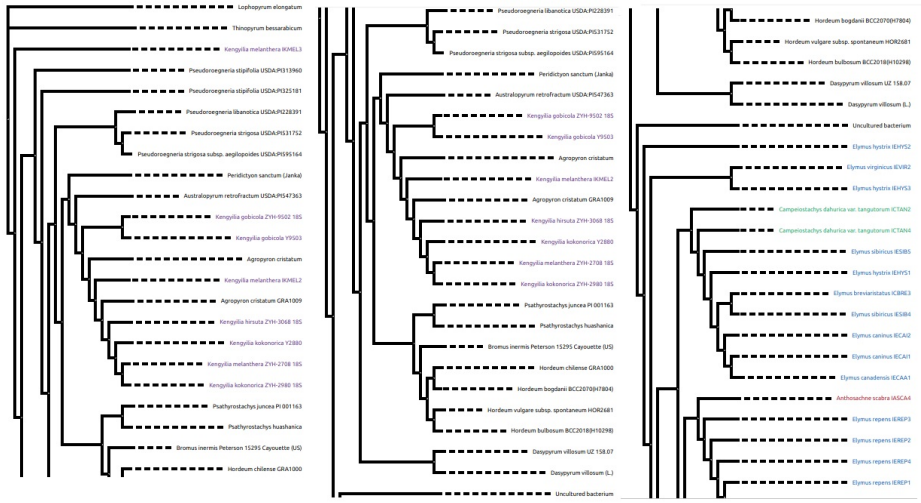

Figure 9: Our reproduction of the matK tree, using Ugene software.

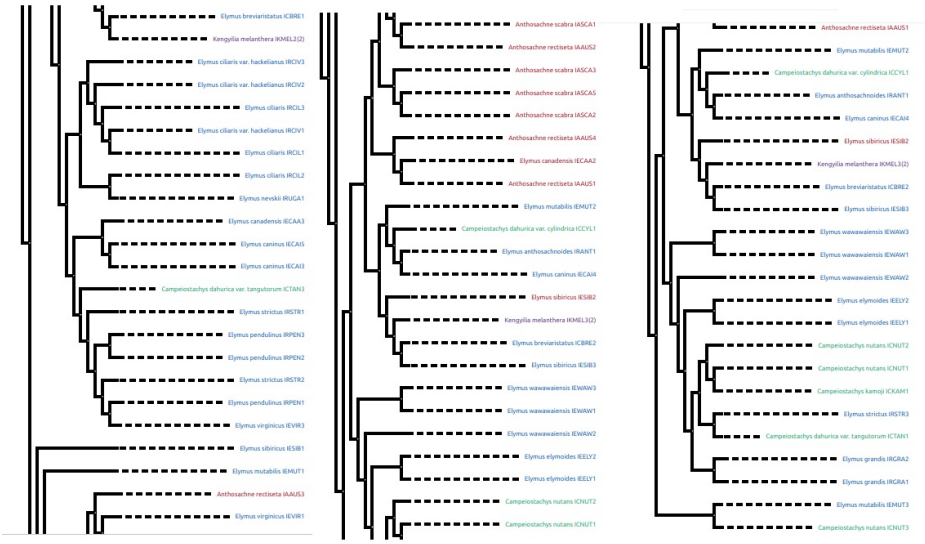

Figure 10: part 2 - Our reproduction of the matK tree, using Ugene software.

- trnH-psbA analysis

In the original article, the ML analysis of the trnH-psbA sequence data. Likelihood settings from best-fit model (K81uf + G) were selected. In their tree for the trnH-psbA, they obtained two different St-type trnHpsbA sequences from *Elymus* s. l. species.

In our replication of the trnH-psbA tree [Figure 11], even if the best model in the original paper was k81uf + G, due to limitations in the software Ugene, was not possible to select it. Thus, we ended up using only the k81uf, a more generic model.
Because of that, the configuration of the tree ended up being different. The clades are still visible, however, in different positions.

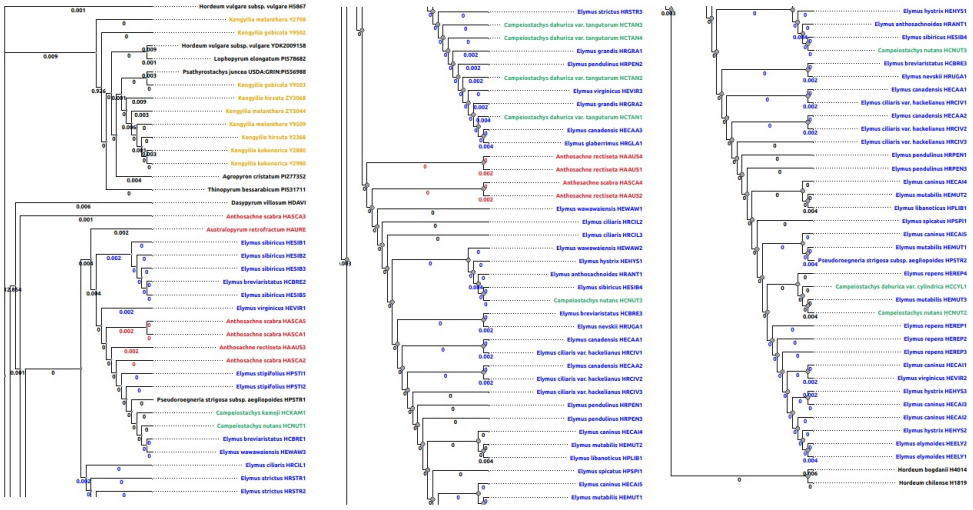

Figure 11: Our reproduction of the trnH-psbA tree, using Ugene software.

4 Conclusion

The extraction of the genome data from NCBI, using our script, was successful, and it depends on the user input in the terminal.
We found a minor inconvenience in the aligning of sequences where, although it was done correctly, had to done in two parts, generating two files for each gene, to be merged in the MEGA software.
And finally, the creation of the phylogenetic trees, we managed to draw them, using Ugene software, and were colored in a way that they are readable and similar to those in the original aricle.
So, overall, we may say that our goals were reached successfully.

References

[1] Dutta, S. S., PhD. (2021, March 9). Phylogenetic Analysis. News-Medical.Net. (Available at news-medical.net/health)

[2] Dong, ZZ., Fan, X., Sha, LN. et al. Phylogeny and differentiation of the St genome in Elymus L. sensu lato (Triticeae; Poaceae) based on one nuclear DNA and two chloroplast genes. BMC Plant Biol 15, 179 (2015). (Available at bmcplantbiol.biomedcentral.com) https://bmcplantbiol.biomedcentral.com/articles/10.1186/s12870-015-0517-2author-information

[3] Zhang, K., Wang, X., Cheng, F. (2019). Plant Polyploidy: Origin, Evolution, and Its Influence on Crop Domestication. Horticultural Plant Journal. (Available at sciencedirect.com)

[4] Goulet, B. E., Roda, F., Hopkins, R. (2017). Hybridization in plants: Old ideas, new techniques. Plant Physiology. (Available at academic.oup.com)

[5] Canonical. (2019, June 19). The leading operating system for PCs, IoT devices, servers and the cloud — Ubuntu. Ubuntu. ubuntu.com

[6] What is Linux? (2018, August 20). Linux.Com. www.linux.com/what-is-linux

[7] Oracle VM VirtualBox (6.1.34). (2022). [Virtualization software]. www.virtualbox.org

[8] GitHub, inc (2022). Docs.Github - quickstart. docs.github.com/en/get-started/quickstart/hello-world

[9] Repository (2022). GitHub - Princesacorderosa/Assignment-ASB: ASB - Research on a phylogenetic analysis paper. Recreation of methods and studies. https://github.com/Princesacorderosa/Assignment-ASB

[10] Introduction to Bash (Shell). (2020, September 14). Earth Data Science - Earth Lab. https://www.earthdatascience.org/courses/intro-to-earth-data-science/open-reproducible-science/bash/

[11] NCBI - National Center for Biotechnology Information. NCBI. www.ncbi.nlm.nih.gov/

[12] MAFFT - Multiple Alignment using Fast Fourier Transform. (2021). [Alignment software]. Kazutaka Katoh. mafft.cbrc.jp/alignment/software

[13] MEGA (Molecular Evolutionary Genetics Analysis). (2022). [Software]. ttps://www.megasoftware.net/

[14] Okonechnikov K, Golosova O, Fursov M, the UGENE team. Unipro UGENE: a unified bioinformatics toolkit . Bioinformatics 2012 28: 1166-1167. doi:10.1093/bioinformatics/bts091 https://ugene.net/

[15] Overleaf and LaTeX Documentation. (2022). Overleaf. overleaf.com/learn