Serdica J. Computing **12** (2018), No 3, 163–174

**Serdica**
Journal of Computing

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

# *IN SILICO* PREDICTION OF $C_4$-RELATED GENES BY FINDING DUPLICATIONS CAUSING PATTERN DEVIATION AND COMPARATIVE ANALYSIS OF PHYLOGENETIC TREES

Irena Avdjieva, Milko Krachunov, Dimitar Vassilev

Abstract. This study is focused on the development of a pattern-finding method for analyzing evolutionary trees to predict genes that may be involved in $C_4$ photosynthesis. It relies on publicly available phylogenetic data which is processed with the authors' own Python scripts and open-source software. The pattern recognition in the topology of the trees is an essential part of the process and the result is then validated by comparing the expression levels of the selected candidates. The same approach can be applied in studying the evolution of other important traits just by changing the type of pattern.

**1. Introduction.** In this study, we use a computational approach to propose a solution for a biological problem that could offer a more detailed understanding of the process of photosynthesis in so-called $C_4$ plants.

*ACM Computing Classification System* (1998): J.3, H.3.3.
*Key words*: computational phylogenetics, pattern recognition, evolutionary tree, homology.

Photosynthesis is the natural method to produce organic compounds and oxygen ($O_2$) by using atmospheric carbon dioxide ($CO_2$), water and sunlight. Only green plants can perform photosynthesis and this ability makes them an essential, and primary, part of the biosphere. There are two major photosynthetic pathways. The most common is called $C_3$ (because the first product is a 3-carbon compound) and is typical for most plants. In the $C_4$ pathway, the first product is a 4-carbon compound, and this modification is currently observed in 3% of known plants. There are important physiological differences associated with $C_3$ and $C_4$ photosynthesis, many of which have an ecological significance [1, 2].

$C_4$ photosynthesis is a subject of great interest in the past few years because it allows plants to minimize water loss and utilize atmospheric $CO_2$ more efficiently in warm and dry conditions. It is an especially important trait in agriculture and predicting the genes involved in this pathway is a key to the development of drought-resistant crops.

Research shows that $C_4$ photosynthesis has evolved independently more than 60 times during the evolution of green plants. There are several hypotheses about the development of this modification: 1) genes that are present in $C_4$ plants but not in $C_3$ plants (or vice versa); 2) genes that are duplicated in $C_4$ plants but are present as single copies in $C_3$ plants (or vice versa), and 3) copy number variations between homologous genes in one of the two groups [3, 4, 5]. When taking into account recent research on $C_4$ plants, the “duplicated vs. single copies” hypothesis is considered most accurate.

To predict genes that may be involved in $C_4$ photosynthesis this study proposes a comparative phylogenetic approach that involves finding a certain pattern in the topology of the trees which contain genes form both $C_3$ and $C_4$ species. Unlike other research groups which rely on sequencing data, some of which is obtained for poorly annotated plant species, our approach (as described in Fig. 1) uses publicly available phylogenetic trees as source data. This saves the need to do sequence analyses and covers a large part of the genomes of the species involved in this study.

**2. Computational challenges concerning phylogenetic data.** The mail goal of this study is to propose a computational method for predicting $C_4$ genes based on the discovery of a certain pattern in the structure of an evolutionary tree (also called a phylogenetic tree, or phylogeny). It is a reconstruction of the evolutionary history of a group of taxa (in this case—genes), and traces the origin of contemporary traits from a common ancestor. The tree assumes the form of a directed acyclic graph.

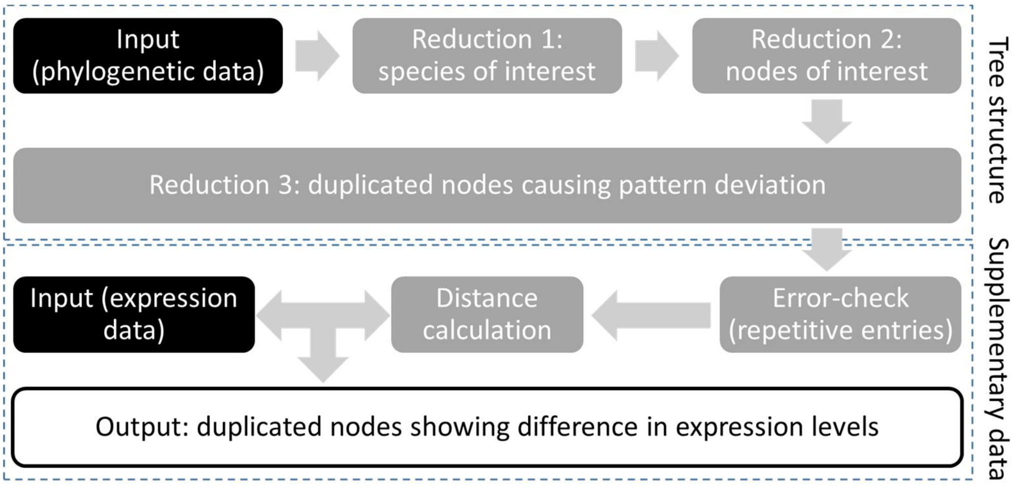

Input (phylogenetic data)
Reduction 1: species of interest
Reduction 2: nodes of interest
Reduction 3: duplicated nodes causing pattern deviation
Tree structure
Input (expression data)
Distance calculation
Error-check (repetitive entries)
Output: duplicated nodes showing difference in expression levels
Supplementary data

Fig. 1. A workflow of the proposed solution: the first part involves the tree structure and includes defining the objects of interest and pattern-finding; the second part validates the results by using additional information and relies on the supplementary data for each node (gene)

In order to find suitable candidate-genes for our study, we started with a large dataset of plant gene trees from different species. We chose to work with publicly available phylogenetic data from the database Ensembl Plants [6]. It contains more than 40 species—mostly model and/or economically important plants. The dataset consists of more than 100 000 gene trees and had to undergo several stages of filtering by various criteria so that less than 100 genes would be proposed as candidates for involvement in $C_4$ photosynthesis (see Fig. 2).

The input data for this research is an EMF (Ensembl Multi Format) flat file dump containing phylogenetic trees in Newick format [7] along with a block of supplementary lines, containing information about each gene in the corresponding tree. The individual entries are separated by two vertical slashes

(//) and the tree structure format was separated from the supplementary block with a single line (DATA). We use information from the supplementary block to filter our data with customized Python 2.7 scripts. For reading and, if necessary, modifying the tree structure we combine our own scripts with the Python-based toolkit E. T. E. [8]—a powerful tool for exploring and analysing phylogenetic trees.

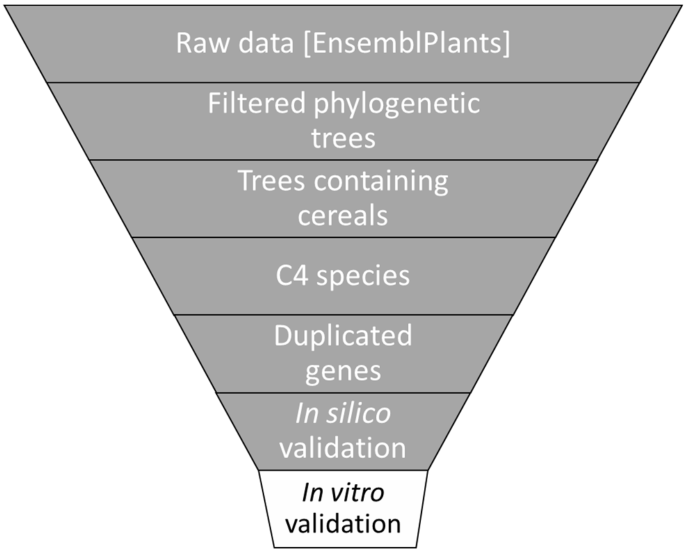


Fig. 2. Step-by step reduction of the dataset

**2.1. Pattern setup.** Before proceeding to search for genes potentially involved in $C_4$ photosynthesis, it was necessary to select appropriate objects. Most $C_4$ plants are grasses, therefore he study was focused on this group and two representatives from the two photosynthetic groups were chosen: $C_3$ plants rice (Oryza sativa) and stiff brome (Brachypodium distachyon), and $C_4$ plants maize (Zea mays) and sorghum (Sorghum bicolor). The reason for choosing exactly these four species to study the evolution of $C_4$ photosynthesis is justified not only by the fact that they belong to the same family, but mainly because they are subject to intense study because of their economic value. According to FAOSTAT [9], rice, maize and sorghum are ranked respectively in first, third and fifth places of world-wide cereals, and brome is a model plant for all grasses [10]. Therefore, their genomes are better annotated than those of a number of other plants.

It is known that the typical ratio between the genes of these four species in terms of evolution is 1: 1: 1: 2 [11]. This means that, typically, in a clade containing these four species, for each rice / brome / sorghum gene there are two maize genes. The reason for this is that maize undergoes a whole genome duplication event dating back to 5-12 million years after the speciation event which led to the separation of maize and sorghum individual species. Thus, a pattern can be observed in the topology of the trees which contain genes from these four species, as shown in the example on Fig. 3. If this pattern has changed and the ratio between genes is no longer 1: 1: 1: 2, the reason behind this deviation is an evolutionary event which has led to gain, loss or duplication of gene(s). In the case of our study, our goal is to search for deviations caused by gene duplication in one or more of the four subjects.

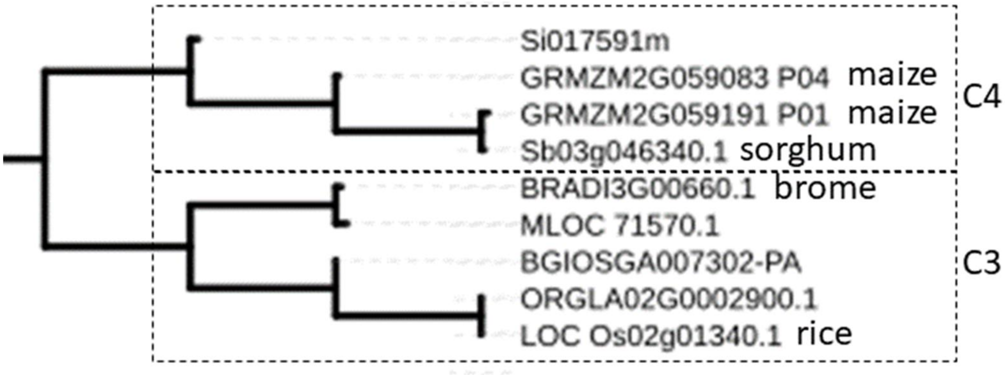


Fig. 3. An example of the topology of a clade (sub-tree) containing genes that match the typical ratio. It is clearly seen that $C_3$ and $C_4$ species have evolved separately, forming two different subtrees, or clades.

**2.2. Dataset preparation and pattern-finding script.** Simply counting the genes from the species of interest and calculating the ratio between them for the entire tree would not work for larger trees containing more than one clade of grasses and may lead to misleading results. Thus, to find deviation from the pattern shown on Fig. 1, we developed a script that reads the main dataset, searches for the smallest clades where all four species are present, and returns a list of the nodes (genes) that do not match the expected ratio, along with their corresponding trees. This task includes two major stages:

1. Defining the objects of interest by the first few letters of the genes' name that correspond to the species name (rice—Os, brome—BRADI, sorghum—Sb, maize—GRM).
2. Defining the expected standard ratio—1 Os: 1 BRADI : 1 Sb: 2 GRM.

```
#set subjects of interest (species name, expected ratio, genes of interest)
define trees of interest:
    for subtree in interest(tree):
        yield subtree
    if gene.startswith(interest):
        yield tree
class GeneInfo(object):
    define (#fields in supplementary lines):
    define from_line(): #obtain GeneInfo by parsing a supplementary line
        fields = line.split()
class TreeInfo(object):
    define (#parts of a whole tree)
    define from_lines(): #obtain TreeInfo by parsing a group of lines
        if line.startswith('SEQ'): #supplementary line
        elif line.startswith('('): #tree structure
        else:#tree index number
        tree = ete2.parser.newick.read_newick(treedata) #use ETE2 to parse
the tree structure
    define pattern-mismatch(self): #return pattern-mismatch subtree
        for subtree in interesting(tree): #calculate ratio between genes
            if ratio = pattern: return False
            else: return subtree
    write in output (supplementary lines of mismatch-causing genes + tree
structure + tree index number)
```

Fig. 4. A simplified description of the pattern-finding algorithm. Comments are marked by a hashtag symbol (#).

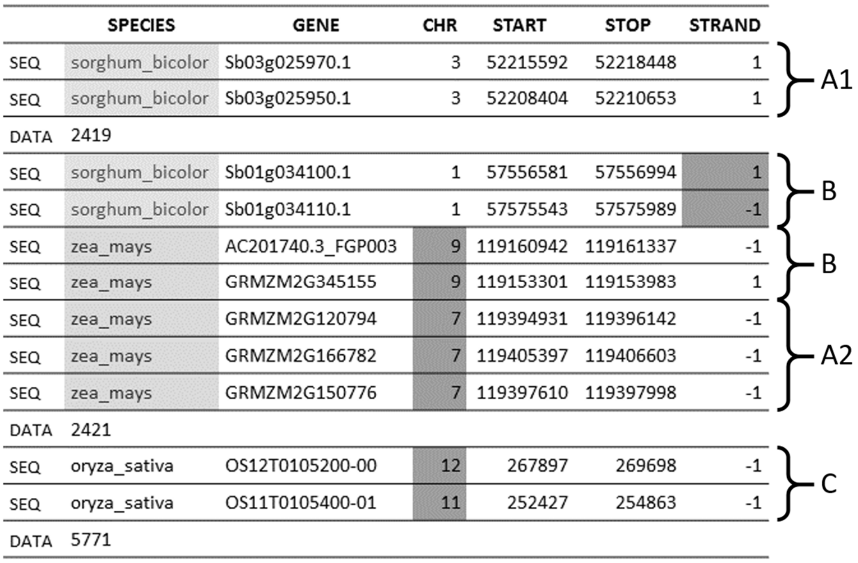
The script (see Fig. 4) reads the tree form leaves to root and finds the smallest clade (subtree) containing all four species. Then, the genes of interest are counted, the ratio between them is calculated and compared to the standard ratio, using TRUE/FALSE statement. If the ratio matches the standard (TRUE), the search continues towards the root of the tree. If there is a mismatch (FALSE), the genes that cause this deviation are recorded with their corresponding supplementary lines in the informative block of the tree. Then the search continues and when the root of the tree is reached, the DATA row, containing the tree index number, the tree structure, and the separator (//) are also recorded.

Due to the script reading a tree from the leaves to the root, some genes can be recorded into the list more than once due to the clades being nested within each other. This requires an extra step to search the output file and remove the repetitive names except the first occurrence, thus leaving only unique genes.

**3. Verification of the results.** For a clearer visualization of the results, the output file is recorded as comma-separated values, which can be read from both a text editor and Microsoft Excel. It is a list of supplementary data for genes that do not match the given ratio, followed by the tree index number (DATA).

The next criterion for further reduction of the list of candidate genes is based on the assumption that the smaller the distance between genes along the chromosome, the more likely they are the product of recent duplication. This task was solved using Microsoft Excel. In order to facilitate the analysis, the information fields were formatted in separate columns as the information is used in the next steps.

Before proceeding to calculate a distance, it is necessary to distinguish individual duplicated groups which may be present in a single tree. When only two or three genes belonging to a species are present within one tree, they can be unambiguously referred to the same group causing the deviation. When the genes are four or more, it is necessary to further specify whether they are part of a single group of duplicated genes or should be considered as separate groups. It is easy to determine which of the two options is involved by checking whether the genes are located in the same chromosome or not. This is accomplished by an *IF* statement, which checks whether the consecutive matching names in the *Species* column match the contents of the *Chromosome* column for the corresponding genes. Possible options are illustrated by the examples given in Fig. 5. The first group is a valid pair of duplicated genes—same species, same chromosome, same DNA strand. The second group shows genes from one species (Sb), but they are located in different DNA strands, which is an error in the dataset and is not considered a duplication. Then there are five genes from one species (GRM) containing two groups of duplicated genes, located in two chromosomes (CHR). The first group has the same error in DNA strands, and the second is a valid triplication. The last group is another error, showing genes located on different chromosomes.



| | SPECIES | GENE | CHR | START | STOP | STRAND |
|---|---|---|---|---|---|---|
| SEQ | sorghum_bicolor | Sb03g025970.1 | 3 | 52215592 | 52218448 | 1 |
| SEQ | sorghum_bicolor | Sb03g025950.1 | 3 | 52208404 | 52210653 | 1 |
| DATA | 2419 | | | | | |
| SEQ | sorghum_bicolor | Sb01g034100.1 | 1 | 57556581 | 57556994 | 1 |
| SEQ | sorghum_bicolor | Sb01g034110.1 | 1 | 57575543 | 57575989 | -1 |
| SEQ | zea_mays | AC201740.3_FGP003 | 9 | 119160942 | 119161337 | -1 |
| SEQ | zea_mays | GRMZM2G345155 | 9 | 119153301 | 119153983 | 1 |
| SEQ | zea_mays | GRMZM2G120794 | 7 | 119394931 | 119396142 | -1 |
| SEQ | zea_mays | GRMZM2G166782 | 7 | 119405397 | 119406603 | -1 |
| SEQ | zea_mays | GRMZM2G150776 | 7 | 119397610 | 119397998 | -1 |
| DATA | 2421 | | | | | |
| SEQ | oryza_sativa | OS12T0105200-00 | 12 | 267897 | 269698 | -1 |
| SEQ | oryza_sativa | OS11T0105400-01 | 11 | 252427 | 254863 | -1 |
| DATA | 5771 | | | | | |

Fig. 5. Examples illustrating possible variants for duplicate genes.
A – valid genes (A1 – duplication, A2 – triplication);
B – error, different strands; C – error, different chromosomes.

As can be seen from the figure, errors are reported as follows:

- Location of the genes in the genome—the duplicated genes are located on the same chromosome, which can be checked in the Chromosome field of the informative part.
- The direction of reading the DNA strand—In order for subsequent analyzes to be performed properly, it is necessary that the entire group of duplicated sequences be oriented in the same direction. This is checked in the Strand field of the informative part.

Once this has been solved, the distance between duplicated genes can be calculated. Information about this can be obtained indirectly from the Start and Stop columns as they contain the start and end positions of each gene along the chromosome. Thus, the distance between duplicated genes is calculated as following:

$$\text{Dist} = \text{start(A2)} - \text{stop(B1)} \text{ for two genes}$$

or

$$\text{Dist} = \max\ (\text{start(A1:A}n)) - \min(\text{stop(B1:B}n)) \text{ for } n \text{ genes}$$

The calculations were then analyzed to show how many trees are retrieved for various distances and the results showed that a plateu is reached at 20 000 base pairs. Only gene groups below this distance have been selected to continue the analysis by comparing the expression levels of the duplicated genes.

**4. Conclusions.** The current in silico approach addressing the evolution of $C_4$ traits relies on finding and tracing a repeatable pattern in the topology of trees containing genes form well annotated $C_3$ and $C_4$ cereals. The results shall be validated by comparing the expression levels of duplicated gene groups—an approach used by other authors in the same field. Additional validation could be carried out by comparing the topology of predicted candidates with that of referent genes whose role in $C_4$ photosynthesis is experimentally confirmed.

This evolutionary approach is an alternative to most other studies on $C_4$ photosynthesis that rely on sequence analyses of a limited number of genes and genomes. The study is entirely based on public datasets which saves both time and resources, and discovers new knowledge in the results of different experiments.

The authors' method for pattern discovery in the topology of phylogenetic trees can be easily modified to address other alternating phenotypes.

**Acknowledgements.** The presented work has been funded by the Bulgarian NSF within the "Methods for Data Analysis and Knowledge Sequencing Datasets" project, Contract DFNI-I02/7 of 12 December 2014.

## REFERENCES

[1] Ehleringer J. R., T. E. Cerling, B. R. Helliker. $C_4$ photosynthesis, atmospheric $CO_2$, and climate. *Oecologia*, **112** (1997), No 3, 285–299.

[2] Buick R. When did oxygenic photosynthesis evolve? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **363** (2008), No 1504, 2731–2743.

[3] Sage R. F. The evolution of $C_4$ photosynthesis. *New Phytologist*, **161** (2004), No 2, 341–370.

[4] Taiz L., E. Zeiger. Plant physiology. Sinauer Associates Inc. Publishers, $4^{\text{th}}$ ed., Massachusetts, 2006.

[5] Kellogg E. A. Phylogenetic aspects of the evolution of $C_4$ photosynthesis. In: R. F. Sage, R. K. Monson (eds). $C_4$ plant biology. Academic Press, 1999, 411–444.

[6] Kersey P. J., J. E. Allen, M. Christensen, P. Davis, L. J. Falin, C. Grabmueller, D. Seth, T. Hughes, J. Humphrey, A. Kerhornou, J. Khobova, N. Langridge, M. D. Mcdowall, U. Maheswari, G. Maslen, M. Nuhn, C. K. Ong, M. Paulini, H. Pedro, I. Toneva, M. A. Tuli, B. Walts, G. Williams, D. Wilson, K. Youens-Clark, M. K. Monaco, J. Stein, X. Wei, D. Ware, D. M. Bolser, K. L. Howe, E. Kulesha, D. Lawson, D. M. Staines. Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Research*, **42** (2014), Database Issue, Article D546–D552.

[7] Archie J., W H. E. Day, W. Maddison, C. Meacham, F. J. Rohlf, D. Swofford, J. Felsenstein. The Newick tree format. 1986. http://evolution.genetics.washington.edu/phylip/newicktree.html, 15 November 2018.

[8] Huerta-Cepas J., J. Dopazo, T. Gabaldón. ETE: a python environment for tree exploration. *BMC Bioinformatics*, **11** (2010), 24.

[9] FAOSTAT. Statistical databases. Food and Agriculture Organization of the United Nations. `http://www.fao.org/faostat/en`, 15 November 2018.

[10] BRKLJACIC J., E. GROTEWOLD, R. SCHOLL, T. MOCKLER, D. F. GARVIN, P. VAIN, T. BRUTNELL, R. SIBOUT, M. BEVAN, H. BUDAK, A. L. CAICEDO, C. GAO, Y. GU, S. P. HAZEN, B. F. HOLT, III, S.-Y. HONG, M. JORDAN, A. J. MANZANEDA, T. MITCHELL-OLDS, K. MOCHIDA, L. A. J. MUR, C.-M. PARK, J. SEDBROOK, M. WATT, S. J. ZHENG, J. P. VOGEL. Brachypodium as a Model for the Grasses: Today and the Future. *Plant Physiology*, **157** (2011), No 1, 3–13.

[11] SWIGOŇOVÁ Z., J. LAI, J. MA, W. RAMAKRISHNA, V. LLACA, J. L. BENNETZEN, J. MESSING. Close split of sorghum and maize genome progenitors. *Genome research*, **14** (2004), No 10a, 1916–1923.

*Irena Avdjieva*
*e-mail:* `i.y.avdjieva@fmi.uni-sofia.bg`
*Milko Krachunov*
*e-mail:* `milkok@fmi.uni-sofia.bg`
*Dimitar Vassilev*
*e-mail:* `dimitar.vassilev@fmi.uni-sofia.bg`
*Faculty of Mathematics and Informatics*
*St. Kliment Ohridski University of Sofia*
*5, James Bourchier Blvd*
*1164 Sofia, Bulgaria*

*Received November 10, 2017*
*Final Accepted November 9, 2018*