A census analysis of the 5-enolpyruvylshikimate-3-phosphate (EPSP) synthase and EPSP-associated domains

Tuomas Tall

Master's thesis

University of Turku
Department of Biology
07.08.2020

Field: Physiology and Genetics
Specialization: Genetics

Credits: 40 ECTS

Reviewers:

1:

2:

Accepted on:

Grade:

UNIVERSITY OF TURKU
Department of Biology
Tuomas Tall

A census analysis of the 5-enolpyruvylshikimate-3-phosphate (EPSP) synthase and EPSP-associated domains

Thesis, 46 pages (8 appendices).

Biology

August 2020

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin Originality Check service.

Background: Glyphosate is one of the most used herbicides against weeds that targets the enzyme 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS). EPSPS is the central enzyme in the shikimate pathway to synthesize 3 essential amino acids in plants, fungi, and prokaryotes. Although this pathway is not found in animals, herbicide may affect the biodiversity of environmental and host-associated microorganisms.

Aims: In this master thesis I will survey the distribution of the EPSPS enzyme in thousands of microorganisms and I will analyse the evolution of the multi domain structure of the EPSPS enzyme in fungi.

Methods: Data was gathered from public databases of proteins (e.g., Pfam and COG). The analysis of the distribution of the EPSPS was performed using Excel functions and a bipartite network was analysed with the program Cytoscape. The Count program was used to assess evolutionary scenarios by Dollo's maximum parsimony, and the phylogenetic trees were visualized with iTOL.

Results: The EPSPS enzyme is widely distributed in archaea, bacteria, plants, and fungi. The multi domain structure of the EPSPS in fungi is strongly associated with six other genes of the shikimate pathway. The most common multi domain structure is composed by a group of five enzymes (HQ synthase, EPSPS, SKI, DHquinase I and Shikimate DH), which I call in this thesis the "Major 5".

Conclusions: The EPSPS multi domain structure in fungi ranges between two to eight domains. The evolutionary analysis shows that the ancestral of fungi had a multi domain structure of six domains. Thus, there have been domain gains and losses throughout the evolution of the EPSPS in fungi. Further investigations are needed to determine the effect of the EPSPS-associated domains to glyphosate resistance. A scientific article that includes data from this master thesis is publicly available as a preprint at biorxiv and submitted to a peer-reviewed Nature Methods journal (appendix VII).

Keywords: Glyphosate, EPSPS, Shikimate Pathway, Fungi, Multi Domain, Bipartite graph, Phylogenetic tree

TURUN YLIOPISTO

Biologian laitos

Tuomas Tall

5-enolipyruvyylishikimaatti-3-fosfaatti syntaasi (EPSPS) ja EPSPS:n liittyvien domainien meta-analyysi

Tutkielma, 46 sivua (8 liitettä).

Biologia

Kuukausi, 2020

Taustatiedot: Glyfosaatti on yleisesti käytetty kasvimyrkky, joka hyökkää 5-enolipyruvyylishikimaatti-3-fosfaattisyntaasin kimppuun. EPSPS on keskeinen entsyymi kasvien, sienten ja mikrobien aminohappojen synteesiin johtavassa sikimaattireitissä. Vaikka sikimaattireittiä ei löydy eläimistä, kasvimyrkky voi vaikuttaa biodiversiteettiin ja hyödyllisiin mikrobeihin.

Päämäärät: Tässä pro gradututkielmassa analysoin EPSPS entsyymin levinneisyyttä mikro-organismeissa ja tutkin sen geneettisen arkkitehtuurin evoluutiota sienissä.

Menetelmät: Data kerättiin julkisista tietokannoista (Pfam, COG). EPSPS geenin levinneisyyden analyysi tehtiin Excelillä ja kaksipuolinen kuvaaja laadittiin ja analysoitiin Cytoscape ohjelmalla. Fylogeneettinen sukupuu luotiin iTOL sivustolla ja COUNT ohjelmalla arvioitiin erilaiset evoluutio skenaariot perustuen Dollon suurinta mahdollista todennäköisyyteen.

Tulokset: EPSPS entsyymi löytyi arkeista, bakteereista, kasveista ja sienistä. Sienissä EPSPS esiintyy monidomeeni rakenteena, jossa esiintyy yleensä 5 muuta sikimaatireitn entsyymiä. Yleisintä monidomeeni rakennetta, mihin kuuluu 6 domeenia: HQ synthase, EPSPS, SKI, DHquinase I ja Shikimate DH, kutsun tutkielmassa Major 5:si.

Johtopäätelmät: EPSPS monidomeenin rakenne vaihtelee sienissä 2–8 domeenin välillä. Evoluutiopuun analyysi paljastaa, että muinaisin monidomeeni on 6 domeenin pituinen. Evoluution aikana monidomeeni on menettänyt ja saanut uusia domeeneja. Lisätutkimusta tarvitaan selvittämään erilaista monidomeenien vaikutuksen sienten glyfosaattivasteeseen. Tieteellinen artikkeli, joka sisältää dataa tästä gradusta, löytyy julkisena koevedoksena Biorxiv sivulta ja se lähetetään vertaisarvioitavaksi Nature Methods julkaisuun (Liite VII).

Asiasanat: Glyfosaatti, EPSPS, Sikimaattireitti, Sienet, Bipartiitti kuvaaja, Fylogeneettinen puu.

CONTENTS

1. INTRODUCTION

1.1. Glyphosate and RoundUP

RoundUp is one of the most common herbicides in the world and it is highly effective against all kind of weeds (Soumis et al. 2018). This herbicide's main active ingredient is isopropylamine salt of a N-(phosphonomethyl)glycine or Glyphosate. Its chemical formula is $C_3H_8NO_5P$ and structure formula is $HOOCCH_2NHCH_2PO(OH)_2$. In 1970 a chemist of Monsanto Company, John E. Franz, found that glyphosate was a herbicide (Franz 1974). Four years later, Monsanto brought the RoundUp herbicide to markets for agricultural use. Glyphosate is classified as a post-emergent non-selective herbicide, which means its effective after germination and throughout plant's growth, and that it kills any vegetation it is used on (Soumis et al. 2018).

Soon after the introduction of RoundUp farmers quickly adopted the new herbicide for weed control. However, as this glyphosate-based herbicide is non-selective, it also affected their crops. In 1996 Monsanto introduced glyphosate-resistant RoundUp Ready crops (e.g. corn, soy, and cotton) so farmers could use company's herbicide without fear of damaging their harvest (Duke 2018). The United States Environmental Protection Agency (US EPA) estimated that between 2005 and 2012, glyphosate-based herbicides (GBH) were the most used herbicide in United States agricultural sector and the second most used (after the herbicide 2,4-D) for non-agricultural usage (Atwood and Paisley-Jones 2017). In 2009 herbicide sales made 10% of Monsanto's revenue and altogether RoundUp line of products made 50% of company's revenue (Cavallaro 2009). In 2018 Monsanto was acquired by the pharmaceutical company Bayer AG (Naomi 2018; Myers et al. 2016; Atwood and Paisley-Jones 2017).

Since 1970, the use of glyphosate based herbicides has increased 100-fold worldwide (Myers et al. 2016; Vandenberg et al. 2017). In parallel to glyphosate usage, there has been an increase of glyphosate resistant weeds, and subsequent concerns about its effects on ecological networks, including (in)direct effects on animals, bacteria and fungi (Richards et al. 2006; Bentley 1990). Moreover, some health organisations such as the International Agency for Research on Cancer (IARC) have warned of potential carcinogenic effects of glyphosate (Guyton et al. 2015; IARC 2015; Vandenberg et al. 2017; Myers et al. 2016).

1.2. Glyphosates' target enzyme: EPSP Synthase

Glyphosate-based herbicides interferes with the shikimate pathway that produces three essential aromatic amino acids (phenylalanine, tyrosine and tryptophan) (Steinrücken and Amrhein 1980; Bentley 1990; Maeda and Dudareva 2012) (Figure 1). The target of this herbicide is the enzyme 5-enolpyruvylshikimate 3-phosphate synthase (EPSPS) and the glyphosate acts as a competitive inhibitor of the second substrate of the enzyme (phosphoenolpyruvate; PEP) —i.e., Glyphosate blocks the binding of PEP with the active site of the EPSPS. The EPSPS enzyme is present in prokaryotes (archaea and bacteria), fungi and plants (Maeda and Dudareva 2012; Funke et al. 2006; Schönbrunn et al. 2001).

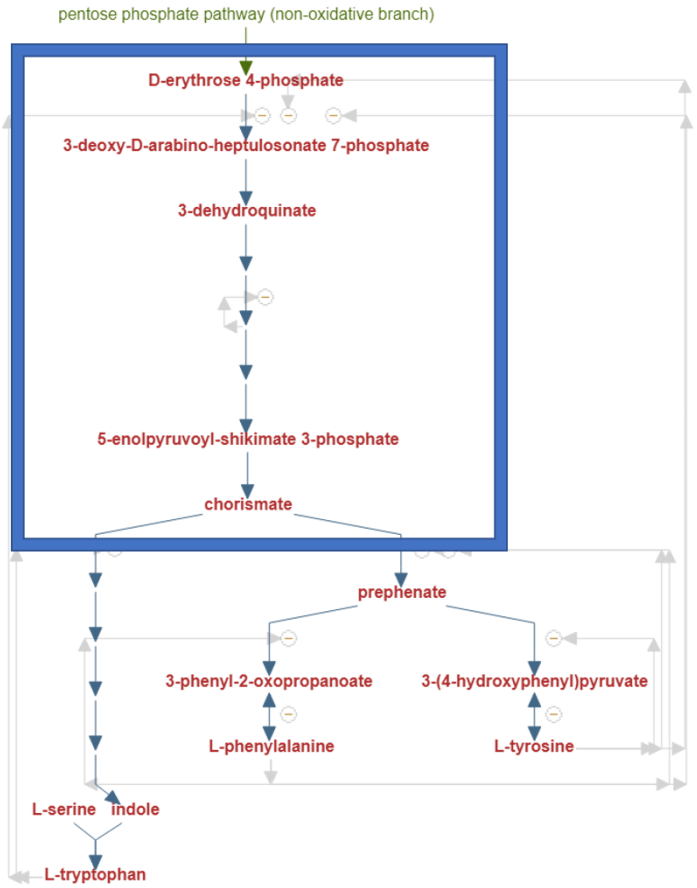

Figure 1. Biosynthesis of tryptophan, phenylalanine, and tyrosine in *Escherichia coli* K-12 substr. MG1655. Blue arrows are different reactions and red text shoves products of reactions. Shikimate pathway is within blue frame. Grey arrows show negative feedback loop where products inhibit the reactions (full details in appendix I; source: https://metacyc.org/ 14.03.2020)

The EPSPS enzyme is not found in animals (with very few exception (Starcevic et al. 2008)) —i.e., the three essential aromatic amino acid (phenylalanine, tyrosine and tryptophan), which are needed for protein synthesis, cannot be synthesized from scratch

and need to be obtained from their diet (Young 1994). The list of essential amino acids varies across species —e.g., humans are not able to synthesize the three aromatic amino acids mentioned plus valine, threonine, methionine, leucine, isoleucine, lysine, and histidine (Maeda and Dudareva 2012; Starcevic et al. 2008; Funke et al. 2006; Herrmann and Weaver 1999).

The EPSPS enzyme catalyses a transference reaction where the substrates phosphoenolpyruvate (PEP) and shikimate-3-phosphate (S3P) are turned into EPSP and phosphate (figure 2) (Jaworski 1972). Substrates go through acetal-like tetrahedral intermediate stage where hydroxyl group in PEP is deprotonated and the enolpyruvate moiety is transferred onto S3P hydroxyl group. As stated before, glyphosate inhibits this reaction by binding to the PEP binding site. In other words, it mimics the intermediate state of the substrate-enzyme complex (Schönbrunn et al. 2001).

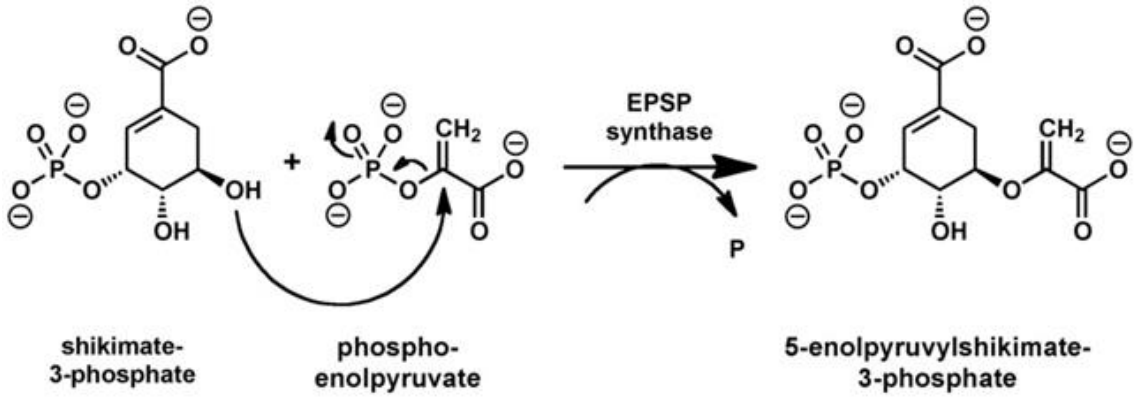

Figure 2. Chemical reaction catalysed by 5-enolpyruvylshikimate-3-phosphate (EPSP) synthase. Source: https://en.wikipedia.org/wiki/EPSP_synthase (CC BY-SA 4.0)

The shikimate pathway is present in prokaryotes (bacteria and archaea) and eukaryotes (mostly plants, algae, fungi, and some protozoan). The protozoan Starlet Sea Anemone (*Nematostella vectensis*) obtained genes of the shikimate pathway *via* horizontal gene transfer from bacteria (Starcevic et al. 2008). Bacterial homologs of shikimate pathway genes (aroA = EPSPS; aroB = 3-dehydroquinate synthase; aroC = chorismate synthase; and aroE = shikimate dehydrogenase) in *N. vectensis* are suspected to come from its bacterial endosymbiont (*Tenacibaculum sp.* MED152). The genome of the endosymbiont is slowly assimilated by host anemone through gene transfer as a mechanism of metabolic

adaptation. As a result *N. vectensis* is most likely cable of producing its own aromatic amino acid for the biosynthesis of polypeptides (Starcevic et al. 2008).

1.3. EPSPS and glyphosates effects of health of organisms

Due to the fact that EPSPS enzyme is not found in animals, glyphosate was considered to be a safe herbicide for general use (EFSA 2015; Tarazona et al. 2017; Boobies 2016; Bundesinstitut für Risikobewertung 2015), however research on long term effects of the herbicide on human is still ongoing (Vandenberg et al. 2017; Tarazona et al. 2017). There have been many reports of toxicity of glyphosate from health organisations, though they are conflicting. The World Health Organisation (WHO) and the International Association for Research on Cancer (IARC) classified glyphosate as category 2A (probably carcinogenic in humans) (WHO 2019; IARC 2015; Guyton et al. 2015) whereas the European Food Safety Authority (EFSA) did not consider glyphosate carcinogenic on itself (EFSA 2015; Europian Union 2015). The German Federal Institute for Risk Assessment 2013 toxicology review did not find clear link between exposure to glyphosate products and non-Hodgkin lymphoma (NHL), while meta-analysis in 2014 by Leah Schinasi and Maria E. Leon did find out that workers exposed to the herbicide have higher risk to gain NHL (Schinasi and Leon 2014). The current consensus from health organisations, such as the WHO, the EFSA, European Chemical Agency (ECHA) and the Bundesinstitut für Risikobewertung (BfR), is that there are no strong evidences to associate glyphosate to cancer (Tarazona et al. 2017; Boobies 2016; Bundesinstitut für Risikobewertung 2015; Mesnage and Antoniou 2017; Clausing et al. 2018). However, some RoundUp additives (e.g., surfactants that facilitate the penetration of the herbicide into the plant cuticle) may be toxic and carcinogenic (Giesy 2000; Mesnage and Antoniou 2017). Some studies have criticised the studies on glyphosate carcinogenesis, pointing out that they used outdated data (Vandenberg et al. 2017; Clausing et al. 2018). An excessive use of glyphosate, among other herbicides, has been linked to the decline of monarch butterfly and bee populations (Pleasants and Oberhauser 2013; Balbuena et al. 2015). Even though humans do not have the EPSPS enzyme, the genome of many gut bacteria have at least one copy of the gene (for the synthesis of three essential aromatic amino acids (Clarke et al. 2014; Cerdeira and Duke 2006)) and may be affected by Glyphosate (Schönbrunn et al. 2001) .

Monsanto has produced genetically modified (GM) crops, resistant to RoundUp, through the insertion of the *EPSPS* gene from *Agrobacterium sp.* strain CP4 (Funke et al. 2006). Thus, as the use of RoundUp in agriculture has increased, so has the number of resistant weeds. In 2014, researchers have found 23 species of glyphosate resistant weeds, which evolved after intense selection pressure in the form of repeated use of glyphosate (Green and Owen 2011). Monsanto researchers have found species that have up to 160 copies of the *EPSPS* gene in their genome (Myers et al. 2016; Vandenberg et al. 2017; Soumis et al. 2018).

1.4. EPSPS protein

The EPSPS is a monomeric enzyme with a molecular mass of 46,000 u. The structure of the enzyme is composed by two subunits, which are joined by protein strands to bring the two protein domains closer together (Sutton et al. 2016). During the process of substrate-enzyme binding, ligand bonding causes the two parts of the enzyme clamp down around the substrates in the active site (Sutton et al. 2016).

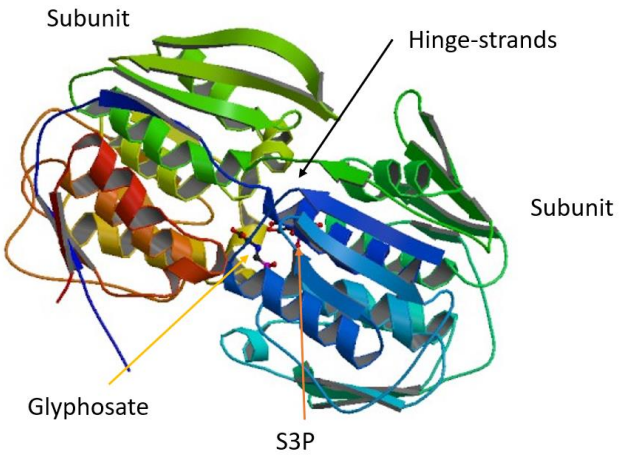

Figure 3. Structure of CP4 EPSP synthase liganded with S3P and glyphosate. Different folds of protein are color coded. From Agrobacterium sp. (strain CP4) by Schonbrunn, E. and Funke, T. on 13.7.2011. Based on X-ray diffraction at 1,70 Å resolution. This strain is resistant of glyphosate and was used in RoundUp Ready GMO crops. Image modified from http://pfam.xfam.org/structure/2GGA on 29.4.2020.

In general, the EPSPS is divided into two classes based on the sensitivity to glyphosate: (1) class I is inhibited by the herbicide and (2) class II is putatively resistant (Light et al. 2016; Firdous et al. 2018; Priestman et al. 2005; Barry et al. 1997). Genetically modified crops, resistant to Roundup, are made with the Class II *CP4 EPSPS* gene from *Agrobacterium sp.* strain CP4 (Funke et al. 2006). Moreover, resistance may be naturally

acquired by means of mutations in the *EPSPS* gene, by gaining multiple copies of the *EPSPS* gene to increase volume of synthesis, or by physiological alterations to reduce herbicide intake to the plant (Pollegioni et al. 2011).

In fungi, EPSPS forms part of a large multidomain protein. Multidomains form when genes code large proteins that have multiple parts or functional units within the protein. Some domains are conserved, but others are highly promiscuous (Basu et al. 2008), i.e. domains capable to combine in multiple protein structures (Basu et al. 2008). Usually duplication is the main mechanism to produce promiscuous domains that play an important role in Protein-Protein Interaction (PPI) networks (Basu et al. 2008; Barrera et al. 2014). An example in the shikimate pathway is the AROM complex. This complex is a pentafunctional polypeptide (polyenzyme with five reactions), which contains enzymes that catalyse steps from two to six of the shikimate pathway (subunits are homologs to separate enzymes) (Lumsden and Coggins 1977b; Hawkins et al. 1993). The structural genes of these five enzyme are found in the Fungi: *Neurospora crassa* (Lumsden and Coggins 1977b), *Aspergillus nidulans* (Moore et al. 1994), and *Saccharomyces cerevisiae* (Duncan et al. 1987). In bacteria, *Bacillus subtilis* has a trifunctional enzyme complex that contains Dehydroquinate synthase, chorismate synthase and NADPH flavin reductase (Hasan and Nester 1978). Another examples of multi-modular enzymes are the non-ribosomal peptide synthetases (NRPS) and the polyketide synthases (PKS), which are responsible of the synthesis of multiple biologically active products produced by bacteria and fungi (Duncan et al. 1987) (Lumsden and Coggins 1977a). Proteins that have the same domain architecture are likely to have similar structures and similar cellular function (Barrera et al. 2014). Moreover, the amount of domain architectures is related to species complexity and lifestyle. In fungi, 10.9% of domains are exclusively found in multidomain proteins, 18.3% of domains are exclusive of single-domain proteins and 70.8% can be found in either single-domain and multidomain proteins (Barrera et al. 2014; Basu et al. 2008).

1.5. Aims of the thesis

The aim of this master's thesis is to perform a census analysis of the domain 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS) and EPSPS-associated domains. This master thesis, which is part of a larger research project to understand the effect of glyphosate on organisms, uses techniques in computational biology such as bipartite

networks, maximum parsimony, and phylogenetic trees. The thesis is divided in three parts:

1) To perform a functional and taxonomical characterization of the *EPSPS* gene in bacteria, archaea, and eukaryotes (mainly plants and fungi).
2) To analyse multidomain organization and diversity of the EPSPS-associated domains in fungi.
3) To study the evolutionary history of domain architectures of the EPSPS-associated domains.

2. MATERIALS AND METHODS

2.1. Data collection

Protein domains data was obtained from a widely used bioinformatics database of conserved domains, called PFAM (Finn et al. 2008) (Sammut et al. 2008) (El-Gebali et al. 2019) (http://pfam.xfam.org). Information of the EPSPS-associated domains from public repositories, such as PFAM and NCBI (http://ncbi.nlm.nih.gov), was provided by Dr Puigbò in excel format. I also used data from an NCBI database of clusters of orthologous groups to study spread of the EPSPS protein in prokaryotes (Galperin et al. 2015; Tatusov et al. 2000) (https://www.ncbi.nlm.nih.gov/COG). Altogether, the combined dataset was used for the analysis of domains through a comprehensive census analysis of *EPSPS* genes and proteins.

2.1.1. Protein Families database (PFAM)

A protein family is a group of evolutionarily related proteins. Proteins in a family descend from a common ancestor and typically have similar 3D structures, functions, and significant sequence similarity (The European Bioinformation Institute and The European Molecular Biology Laboratory n.d.). Proteins are generally composed of one or more functional or structural region, called domains. They are independent stable tertiary structures that can evolve, exist, and function separately. Different combinations of domains give rise to the diverse range of proteins found in nature. The identification of domains that occur within proteins provide insights into proteins function as well as to study the evolutionary history of proteins (Barrera et al. 2014; Basu et al. 2008). The Pfam database provides a comprehensive collection and classification of protein families and domains, each represented by multiple sequence alignments (MSA) with hidden Markov models (HMMs) (Sonnhammer et al. 1997). The Pfam database (version 32.0)

was used to gather taxonomic information and multidomain architectures. Instead of using BLAST (basic local alignment search tool) algorithm for search, Pfam uses profile HMMs algorithm, which give greater weight to matches at conservative sites by taking account gap and substitution probabilities which better reflect biological reality (Finn et al. 2008; Xu and Dunbrack 2012; Sammut et al. 2008; Sonnhammer et al. 1997; Altschul et al. 1990; Eddy 1998). Profile HMMs are probabilistic models that encapsulate the evolutionary changes that have occurred in a set of related sequences (i.e. a MSA) (Bishop and Thompson 1986; Stratonovich 1960). The MSA of homologous (evolutionary relationship between sequences that are descended from a common ancestor) sequences is a preliminary step to determine their evolutionary history (Sobel and Martinez 1986).

2.1.2. Clusters of Orthologous groups (COG)

Homology was defined original in 1843 by Sir Richard Owen as "the same organ under every variety of form and function" (Fitch 2000), and adapted later to describe the evolution of gene and genomes. Orthologs are homologous proteins that descended from the same ancestral protein sequence separated by a speciation event (Tatusov et al. 1997; Galperin et al. 2015). In this thesis I use the database of Clusters of Orthologous Groups (COG) (Galperin et al. 2015) to study the EPSPS (COG0128) presence in prokaryotes (archaea and bacteria). COGs are generated by comparing known and predicted protein sequences from all completed microbial genomes to infer sets of orthologs (Tatusov et al. 2000). Each cluster consists of proteins found to be orthologous at least in three lineages and likely corresponds to an ancient conserved protein. The COGs are constructed by applying the criterion of consistency of genome-specific best BLAST (Altschul et al. 1990) hits, to the results of an exhaustive comparison of all protein sequences from these genomes. COGs are widely used to study evolutionary relationship of proteins, genomes, and species.

2.2. Functional and taxonomic distribution analysis the EPSPS in fungi

First, I gathered information from the literature to understand the current knowledge on the EPSPS domain in fungi. Then, collected data of EPSPS proteins from the PFAM database to quantify the frequency of the domain EPSPS and EPSPS-associated domains (http://pfam.xfam.org) (Sonnhammer et al. 1997). This analysis was performed using Excel functions to calculate the number of different domains found in a sample and to characterize them through Pfam links. Afterwards, I divided these domains in to 4

categories and made a Venn's diagram to classify domains into four partially overlapping groups: shikimate (proteins involved in the shikimate pathway), enzymes (proteins with catalytic function), expression (domains whose products are needed in controlling gene expression) and structural function (products that don't have a catalytic function, like binding sites, histones and helix-turn-helix domains). I categorized the data at kingdom and phylum level and counted the number of species. I used descriptive statistics plots to analyse the taxonomic distribution of the EPSPS proteins by kingdom and phylum.

2.3. A Bipartite network analysis of species and domains with Cytoscape

A Bipartite Network (a.k.a. Bipartite graph or bigraph (Techopedia 2017)) is a set of nodes composed into two independent sets or vertices, for example vertices A and B, such that no two nodes within the same set are adjacent (Seymour et al. 2016). Each edge is drawn from node to node, i.e. the connection should be able to connect between any vertex in A to any vertex in B, but no line can go to the same set. Bipartite networks are mostly used in modelling relationships between two entire separate classes of object (Seymour et al. 2016). In biology, bigraphs can be used to connect nodes from two large groups, e.g. gene clusters and genomes to find the evolutionary history of viruses (Iranzo, Koonin, et al. 2016) or fungi and multidomains in this thesis. Currently, the focus in biology has moved from the study of individual biological components to the study of complex biological systems and their dynamics at a larger scale, together with the increase of analytical methods and the computer power (Pavlopoulos et al. 2018). Bipartite graphs are practical way to study the processes of microbial life beyond classic taxonomy and customary genomic analyses, especially in the case of plasmids and viruses that do not share core genes that can be used to build a phylogenetic trees (Iranzo, Koonin, et al. 2016) Moreover, bipartite networks can be used to identify biases in transfer of genes and novel connections between biological entities (Corel et al. 2018), (Pavlopoulos et al. 2018), (Iranzo, Koonin, et al. 2016). In this master thesis I use bipartite networks to determine associations between EPSPS-associated domains and fungal species. Given the large sample size (1176 EPSPS proteins and 46 protein domains) and the partition of the data between two groups, the bigraph is the ideal network analysis. In my analysis, the two vertices will be defined as (A) domains and (B) species to determine how the EPSPS-associated domains are scattered across fungal species. I used a spreadsheet to collect

fungal samples to obtain a final dataset of 390 EPSPS protein sequences from fungi and 20 EPSPS-associated domains in fungal.

The program Cytoscape was used to visualize the bipartite network (Shannon et al. 2003) of EPSPS-associated domains and fungal species. The network was color coded based on taxonomic and domain information from Pfam. I discarded the default network layout and selected the layout of bigraph into group attribute layout by taxonomic groups. Thus, automatically grouped them by the taxonomy codes, each order separately, with domains forming own group. The six most frequent domains are relocated the middle of the network and the other domains in the outside for a more clear visualization.

2.4. Maximum Parsimony analysis with Count

Maximum parsimony is a method of phylogenetic analysis that optimizes the tree based on characters methods. Under this criterion a tree minimizes the total number of evolutionary steps required to explain the information assigned on the leaves (Farris 1970; Fitch 1971), i.e. the shortest possible tree that explains the data set is considered best. Thus, the goal of maximum parsimony methods is to minimize the amount of homoplasy, such as convergent evolution, parallel evolution, and evolutionary reversals. Due to there is no algorithm capable of scanning the whole tree space of large trees in a reasonable time, algorithms use heuristic methods. Moreover, the most parsimonious tree will be always an approximation to the "real" tree due to (among other reasons) sometimes it underestimates actual evolutionary changes that has occurred (Felsenstein 1978).

In this master thesis, I use maximum parsimony methods to evaluate multiple scenarios of gains and losses of domains in fungi. I use Dollon parsimony, which is based on Dollon law of irreversibility:

> "*an organism never returns exactly to a former state, even if it finds itself placed in conditions of existence identical to those in which it has previously lived ... it always keeps some trace of the intermediate stages through which it has passed* " (Gould 1970).

In Dollon maximum parsimony, a character is gained only once in a particular lineage and cannot be regained if it is lost. This method simplifies the evolutionary analysis and allows the reconstruction of phylogenetic tree (Marshall et al. 1994; Lin et al. 2016; Rogozin et al. 2006).

In this thesis, I have analysed the dataset of protein domain architectures from Pfam (Finn et al. 2008; El-Gebali et al. 2019; Sonnhammer et al. 1997) with the program Count (Csurös 2010) to identify parsimony scenarios of domain gain and loss. Count is a program for the analysis of evolutionary scenarios from phylogenetic profiles and other numerical characters (Csurös 2010). The program automatically calculates the theoretically most parsimonious phylogenetic tree from the input data (Csurös 2010) .

2.5. Visualization of phylogenetic trees

The phylogenetic tree was visualized with the interactive tree of life web program (iTOL) (Letunic and Bork 2007). This program is freely available through the web site https://itol.embl.de/ and trees can be interactively pruned and re-rooted (Letunic and Bork 2007) . Users can map various types of data onto the tree like genome sizes or protein domain repertoires. (Letunic and Bork 2007; Letunic and Bork 2019). Phylogenetic tree are important part of genetic studies and graphical tools to represent the evolution of genes and species are essential in modern biology.

3. RESULTS AND DISCUSSION

3.1. Characterization of Domains

3.1.1. Taxonomic distribution of the EPSP synthase

The EPSPS enzyme is the central enzyme in the shikimate pathway for the synthesis of the three essential aromatic amino acids (phenylalanine, tyrosine and tryptophan) in plants, algae, fungi and prokaryotes (Steinrücken and Amrhein 1980; Bentley 1990). This pathway metabolic is absent in animals, so these amino acids must be acquired with the diet or produced by the microbiome. The analysis of EPSPS sequences from Pfam database (Sonnhammer et al. 1997) shows a major distribution of the EPSPS protein in prokaryotes (archaea and bacteria) than eukaryotes, 94% and 6% of sequences respectively (figure 2a). However, given that prokaryotes are more abundant in the dataset and that the gene is essential in archaea, bacteria, plants and fungi, this result is no unexpected. The *EPSPS* gene is present on every major bacteria clade, mostly in Proteobacteria (32%) and Firmicutes (28%) (figure 2b).

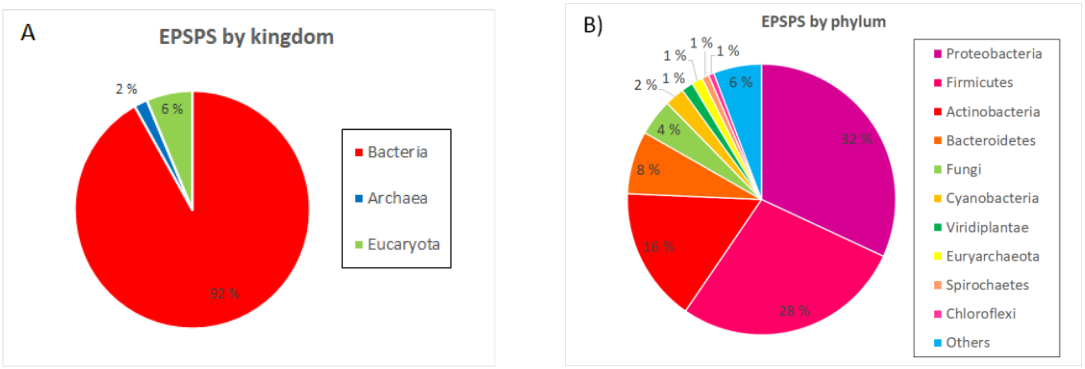

Figure 4. Taxonomic distribution of the *EPSPS* gene: a) by kingdom (N = 9,870). Red is bacterial samples, green eukaryote, and blue archaea. b) by phylum (N = 10,178) Shades of red and yellow are bacterial phyla, greens are eukaryotes (green plants and fungi) and blue is other samples.

The analysis of distribution of the *EPSPS* gene in the COG database (Galperin et al. 2015) (COG0128) shows a widespread of the gene across bacterial taxa, with the exception of parasitic Mollicutes and thermophilic Thermotogae. Bacterial parasites have a tendency to reduce their genome to the minimum by using products from the host (Iranzo, Puigbò, et al. 2016). Nevertheless, other intracellular bacterial parasites have the *EPSPS* gene (e.g., Chlamydia). The absence of the gene in Thermotogae may need further investigation.

In archaea, the *EPSPS* gene is present in most Euryarchaeota, Thaumarchaeota and some species of Crenarchaeota. In eukaryotes it is found in Fungi and Viridiplantae (green algae and plants), Stramenopiles (brown plants) and Haptophyte algae. A BLAST analysis to search for homologous sequences, identifies the *EPSPS* gene in few animals, however these are (most likely) the product of contaminations during sequencing (data source: https://ppuigbo.me/programs/EPSPSClass).

3.1.2. Functional analysis and domain characterization

The EPSPS domain is ~450 amino acids long (~1350 nucleotides) and is present (by definition) in all EPSPS proteins. In most bacteria, archaea and plants, the gene is present as a single domain. However, in most fungi, as well as in bacteria and plants, the gene is part of a multidomain complex. Usually, multidomain *EPSPS* genes of bacteria and plants are formed by pairs of domains together, whereas the fungal EPSPS is a larger sequence

composed of more than five EPSPS-associated domains. The EPSPS-associated domains can be classified in a Venn's diagram into four partially overlapping groups: shikimate (shikimate pathway proteins), enzymes (proteins with catalytic function), expression (domains whose products are needed in controlling gene expression) and structural function (proteins that do not have a catalytic function, like binding sites, histones and helix-turn-helix domains) (figure 5).

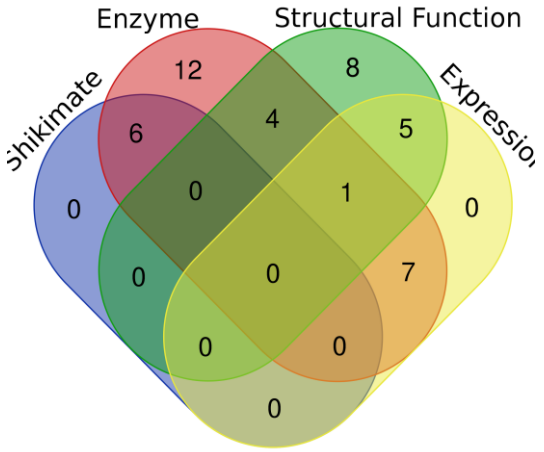

Figure 5. Venn's diagram for the EPSPS domains. Blue is Shikimate pathway, red is enzymes, green is structural function and yellow is gene expression. The diagram was made with a tool from http://bioinformatics.psb.ugent.be/webtools/Venn/.

The EPSPS molecular function is to transfer alkyls aryl (other than methyl) groups. It transfers enolpyruvate from phosphoenolpyruvate to 3-phosphoshikimate. The analysis of the EPSPS-associated domains across species shows that are involved in the Shikimate pathway (SKI, DHQ_synthase, DHquinase_I) and the synthesis of aromatic amino acids synthesis (Shikimate_dh_N, PDH) (Maeda and Dudareva 2012), as well as some promiscuous domains (HTH_3) (Table 1). Least frequency domains are widely distributed across species and theirs functions are DNA modification, gene expression or enzymes. List of all found domains and their frequencies and functions are in Appendix VI.

Table 1. Top 10 most frequent EPSPS-associated domains. Gene: gene names; Freq.: frequency of the gene in the sample; SP: number of species the gene is found; Function: description of product. Rest are found in appendix VI.

Gene	Freq.	SP	Function
EPSPS	1448	8833	6th step of shikimate. pathway, produces EPSP Synthase
SKI	424	7716	Start of pathway, produces phosphorylates shikimate
DHQ_synthase	420	7429	Second step of pathway, removes a phosphate from DHAP
DHquinase_I	416	2073	3rd step of pathway, produces 3-dehydroquinate dehydratase
Shikimate_dh_N	402	7658	The substrate binding domain of shikimate dehydrogenase
HTH_3	218	9670	Helix-turn-helix, a major structural motif capable of binding DNA
Shikimate_DH	160	6719	4th step. Shikimate / quinate 5-dehydrogenase
PDH	127	7136	Prephenate dehydrogenases are part of Tyrosine biosynthesis.
Cytidylate_kin	88	6874	Kinase of cytidine 5'-monophosphate
PF13193	17	8031	AMP-binding enzyme C-terminal domain for PF00501

3.2. Network analysis

3.2.1. Analysis of EPSPS-associated domains in fungi

First, I screened to identify fungal proteins and then I calculated the frequency of domains. A total of 22 domains remained (Table 2) out of 46 domains from the whole dataset (appendix VI). As a control, I verified that the EPSPS domain was in every sample. The most frequent EPSPS-associated domains in 397 EPSPS proteins of fungi are: SKI (Shikimate kinase) (n=374); DHQ synthase (3-dehydroquinate synthase) (n=374); DHquinase (3-dehydroquinate dehydratase) (n=367); Shikimate dh N (Shikimate dehydrogenase substrate binding domain) (n=366); and Shikimate DH (Shikimate / quinate 5-dehydrogenase) (n=136) (Table 2). These domains produce enzymes that are part of shikimate pathway. The least frequent domains were found in only one or two proteins.

Table 2. All domains found in the multidomains of fungal samples. First column is the Pfam code, second column is the name of the gene, third column is the frequency of the domain in the fungal proteins and the fourth column is the function of domain.

Domain	Gene	Freq.	Function
PF00275	EPSP_synthase	397	Produce EPSP Synthase, 6th step
PF01202	SKI	374	5th step of shikimate pathway, phosphorylates shikimate
PF01761	DHQ_synthase	374	Second step of pathway, removes a phosphate from DHAP
PF01487	DHquinase_I	367	3rd step, Hydro-lyase
PF08501	Shikimate_dh_N	366	The substrate binding domain of shikimate dehydrogenase
PF01488	Shikimate_DH	136	4th step. Dehydrogenesis of shikimate to 5-dehydroshikimate
PF05000	RNA_pol_Rpb1_4	2	Domain of RNA polymerase
PF04998	RNA_pol_Rpb1_5	2	Domain of RNA polymerase
PF04997	RNA_pol_Rpb1_1	2	Domain of RNA polymerase
PF04983	RNA_pol_Rpb1_3	2	Domain of RNA polymerase
PF03159	XRN_N	2	5'-3' exonuclease for N-terminus
PF00623	RNA_pol_Rpb1_2	2	Domain of RNA polymerase
PF13932	GIDA_assoc	1	Domain at the C-terminus of protein GidA
PF13716	CRAL_TRIO_2	1	Protein structural domain that binds small lipophilic molecules
PF09320	DUF1977	1	Unknown
PF08241	Methyltransf_11	1	SAM dependent methyltransferases
PF04616	Glyco_hydro_43	1	Glycoside hydrolase family 43, Arabinanases. Found in plants.
PF04098	Rad52_Rad22	1	Double-strand break repair protein
PF01494	FAD_binding_3	1	Monooxygenase, FAD binding in a number of enzymes.
PF01134	GIDA	1	Glucose inhibited division protein A
PF00616	RasGAP	1	All alpha-helical domain that accelerates the GTPase activity of Ras, turning it off.
PF00443	UCH	1	Ubiquitin carboxyl-terminal hydrolase

SKI (Shikimate kinase or ATP:shikimate 3-phosphotransferase) enzyme catalyzes the ATP-dependent phosphorylation of shikimate to form shikimate 3-phosphate. This reaction is the fifth step of the shikimate pathway (Herrmann and Weaver 1999). DHQ synthase (3-dehydroquinate synthase) is quite active carbon-oxygen lyases enzyme that cleaves phosphate from DAHP (3-Deoxy-D-arabino-heptulosonic acid 7-phosphate),

with help of cofactors NAD and cobalt, to produce 3-dehydroquinate.It is the second part of shikimate pathway (Liu et al. 2008; Negron et al. 2011). DHquinase (3-dehydroquinate dehydratase or DHQD) is hydro-lyases that cleaves carbon-oxygen bond from 3-dehydroquinate to produce 3-dehydroshikimate. It is the third part of shikimate pathway (Herrmann 1995). Shikimate DH (Shikimate / quinate 5-dehydrogenase) converts 3-dehydroshikimate to shikimate as well as reduces $NADP^+$ to NADPH. It is the fourth part of shikimate pathway. Shikimate dh N (Shikimate dehydrogenase substrate binding domain) is binding site for the substrate, 3-dehydroshikimate (Ye et al. 2003).

In fungal samples majority (56%) has 5 domains in order of DHQ synthase -> EPSPS -> SKI -> DHquinase I -> Shikimate dh N. When present, Shikimate DH is at the end of sequence after its binding domain. These six domain long sequences make 34% of samples. There are also shorter multidomains, longer ones and ones with alternative domains, but they are minority. Rarer domains are usually at the beginning of multidomain, before SKI. And alternative sequences have either duplication of domains or have lost domains. I refer to the most common five multidomains as Major5 and the six domains as Full6 in this text. These domains are essential genes for production aromatic amino acids. Graph of all found multidomains is in Appendix II.

3.2.2. Cytoscape

A bipartite network, with protein domains and species of fungi forming the nodes connected by edge been based on presence and absence (figure 4A), was constructed with the program Cytoscape (Shannon et al. 2003). The bipartite network connects fungal species divided by Order (ascomycota in the left and basidiomycota, mucoromycota, chytridiomycota, blastocladiomycota, zoopagomycota and unknown in the right). The EPSPS and most common EPSPS-associated domains are located in the centre of the figure (figure 4A) in the same order as in (Figure 4B). Other domains were set on the outside of the bigraph (list of domains in Appendix VI.) and excluded from figure 6A. See more detailed bigraph of separate taxonomic classes of fungi with other domains in Appendix III. Ascomycota seems to be the most variable phylum, in terms of EPSPS-associated domains and contains several rare domains. Moreover, it is the phylum that has least amount of Shikimate DH domains, e.g.: the domain is infrequent in Eurotiomycetes (APE), Dothideomycetes (APD) and Leotiomycetes (APL). Majority of multidomain architectures (n=220) are five domains long, which are involved in the shikimate pathway and ~1/3 of the sequences (n=134) have all six domains (Full 6 domains) of the shikimate pathway. Of those 5 domains long 215 had Major 5 domain

architecture and of 6 domains long 128 had Full 6 architecture (Appendix II). Other 5 and 6 domains long multidomains had duplications of domains, reductions of Full 6 architecture and non-shikimate domains (Appendix II). Small multidomains of 2-4 domains contain reduced number of shikimate domains and some have duplications of the shikimate domains and some non-shikimate domains (n=30). 7 and 8 domains long multidomains have other non-shikimate domains or duplicates (n=7) (Appendix II). Duplication were EPSPS and DHQ Synthase domains (Appendix II). Overall, the bipartite network shows that EPSPS domain in fungi is mostly associated to other domains of the Shikimate pathway. However, there are few fungi that do not the Major 5 or Full 6 domains (e.g., Dothideomycetes). Due to the Shikimate DH domain is found in most species, but not all, it is likely that was present common ancestor of fungi and lost in some lineages. Rarer domains seem to be a later inclusion to multidomain as they appear.

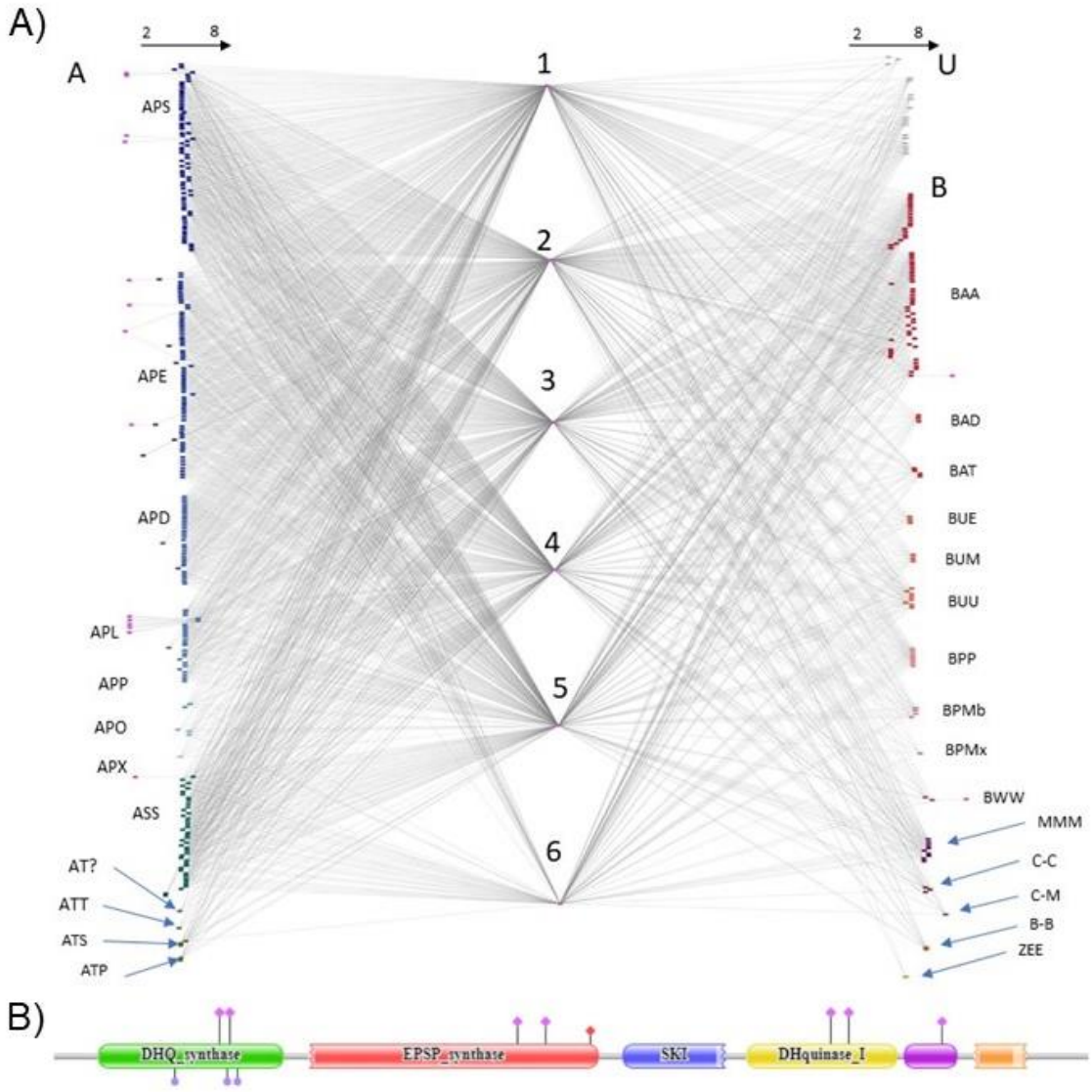

Figure 6. A bipartite network of fungal samples. **A)** A bipartite network of fungal samples. Letters stands for different phyla and classes of fungi. On the left side there are fungi of the phylum Ascomycota (A) and on the right side there are fungi of phylum Basidiomycota (B) Mucoromycota (M), Chytridiomycota (C-), Blastocladiomycota (B-) and Zoopagomygota (Z), Unknown taxa (U). Numbers correspond to fungal multidomain: 1) DH synthase 2) EPSP synthase 3) SKI 4) DH quinase type 1 5) Shikimate dh N and 6) Shikimate DH. Arrows on the top of the figure note the number of domains in multidomain, which ranges between 2 to 8 domains. Subphylum and class starting from upper left are Pezizomycotina: Sordariomycetes (APS), Eurotiomycetes (APE), Dothideomycetes (APD), Leotiomycetes (APL), Pezizomycetes (APP), Orbiliomycetes (APO) and Xylonomycetes (APX); Saccharomycotina: Saccharomycetes (ASS); Taphrinomycotina: incertae sedis (AT?), Taphrinomycetes (ATT), Schizosaccharomycetes (ATS) and Pneumocystidomycetes (ATP); Agariomycotiina: Agaricomycetes (BAA), Dacrymycetes (BAD) and Tremellomycetes (BAT); Ustilaginomycotina: Exobasidiomycetes (BUE), Malasseziomycetes (BUM) and Ustilaginomycetes (BUU); Pucciniomycotina: Pucciniomycetes (BPP), Microbotryomycetes (BPMb) and Mixiomycetes (BPMx); Wallemiomycotina: Wallemiomycetes (BWW); Mucoromycotina: Mucoromycetes (MMM); Chytridiomycota: Chytridiomycetes (C-C) and Monoblepharidomycetes (C-M); Blastocladiomycota: Blastocladiomycetes (B-B); Entomophthoromycotina: Entomophthoromycetes (ZEE). A detailed analysis of the bipartite network can be found in appendix III. **B)** Example of a multidomain protein of 1616 amino acids. The EPSPS and associated domains of the EPSPS protein A0A060S9A7_PYCCI from *Pycnoporus cinnabarinus*. In this example, the SKI and Shikimate DH are fractured.

3.3. Phylogenetic trees and EPSPS-associated domains evolution

3.3.1. Phylogenetic tree

Using ITOL program and data from Pfam we created a phylogenetic tree of the EPSPS domain in fungi (figure 8). This phylogenetic tree is rooted using bacterial outgroups due to it been a subset of a larger tree that includes bacteria, archaea and eukaryotes (appendix V). The inner part of the figure contains the phylogenetic tree with branch length representing evolution time between nodes. Outside of tree are nodes which are each a different sample with names highlighted in red. The colored bars represent the distribution of multidomains (figure 8): DHG synthase (blue), EPSP synthase (green), SKI (yellow), DH quinase type 1 (purple), Shikimate dh N (turquoise) and Shikimate DH (dark red). Other domains (e.g., GIDA, GIDA assoc, XNR) are found in the beginning of multidomains in line of previous analyses. The full tree is freely accessible at https://itol.embl.de/tree/13023210641470501567596434#. The outer ring contains the taxonomical codes used in the bipartite network, which were used to help analyze the phylogeny of tree. From the tree it is clear that Shikimate DH is present across fungal branches. Shikimate DH is absent in some lineages, however but can be found in their sister branches, which suggests recent loss of the domain. Thus, this suggests the hypothesis that shikimate DH is part of original multidomain and has been lost independently (in multiple branches) during the evolution of fungi. Noteworthy is the conservation of the multidomain structure across the phylogenetic trees (figure 8).

Most fungi belong to a monophyletic group within the species tree and that all the species contain multi domains (appendix VI). In the appendix VI there is another monophyletic group of (mostly) plant species separate from the fungal cluster, which suggests either independent horizontal gene transfers from bacteria or an ancient duplication of the gene. The EPSPS domain is present in other eukaryotes spread across the species tree, but some are most likely contaminations (Leino et al. 2020). Overall, the taxonomy of fungi is in constant flux as new DNA analyses upturn old classification that was based on morphology and experimental reproduction. Especially higher taxonomic levels change names frequently (Hibbett et al. 2007). The phylogenetic tree contains phyla (or divisions): asco- (sac fungi), basidio- (club fungi), blastocladio- (Saprotrophs), chytridiomycota (mobile single cell organisms) and subphyla incertae sedis (meaning uncertain placement): mucoro- and entomophromycotina, formerly part of mucoro- and zoopagomycota (Hibbett et al. 2007). This leaves out phyla glomeromycota (arbuscular mycorrhiza symbionts), microsporidia (unicellular parasites) and neocallimastigomycota

(anaerobic fungi without mitochondria) (Hibbett et al. 2007). Samples did not include all classes or species under four phyla. For example, ascomycota samples contained lot of sordariomycetes but no geoglossomycetes or lecanoromycetes class of fungi. Many species of fungi are still unclassified.

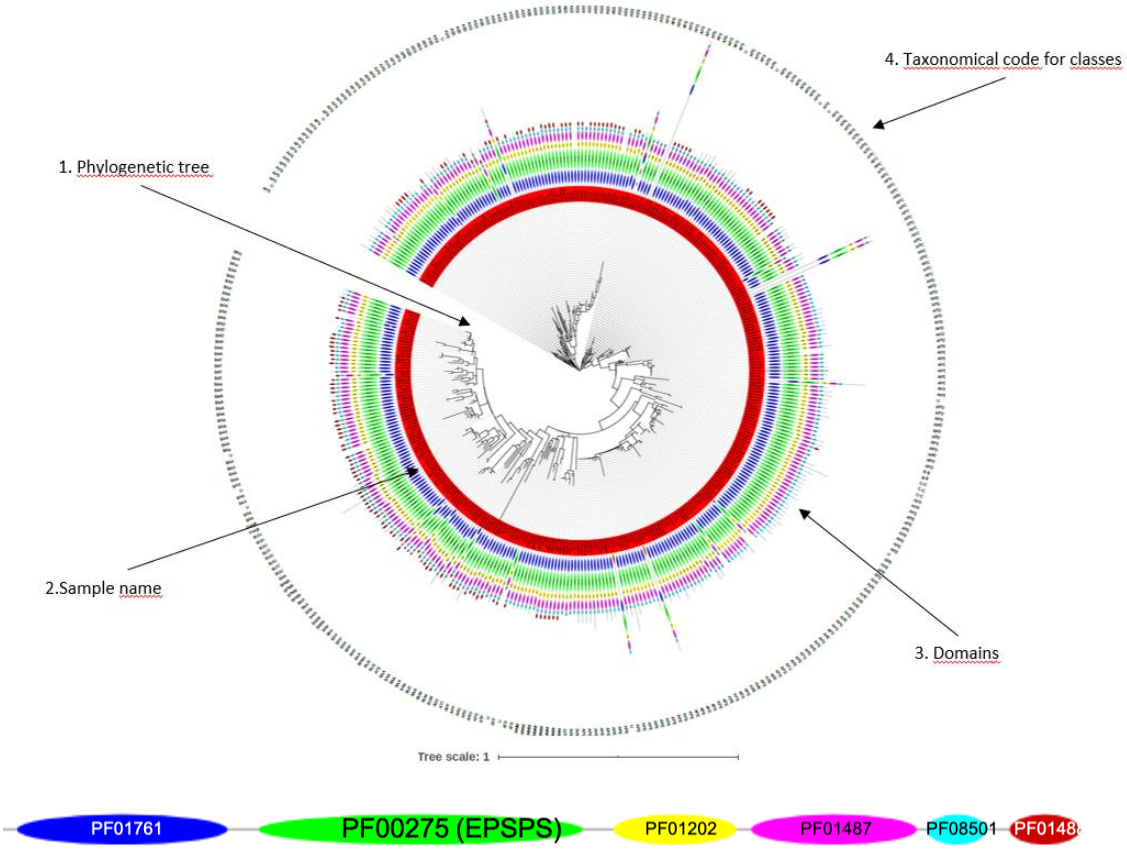

Figure 8. Phylogenetic tree of fungal samples. Centre shows the tree with different branch lengths. Red circle contains names of samples. Outside domains of samples are shown: DHG synthase is in blue, EPSP synthase is in green, SKI in yellow, DH quinase type 1 is purple, Shikimate dh N is turquoise and Shikimate DH in dark red. Rare domains are shown in different colors. This also shoves the whole length of multidomain as grey line. Outer ring is the three-letter code of taxonomic classes used in bigraph.

The phylogenetic tree based on the EPSPS domain (figure 8) does not strictly reflect species tree of fungi based on a current comparison with recent articles (Tedersoo et al. 2018; Silar 2016). Closer look reveals that phyla are mixed together, with blastcladio-, chytridio- and mucoromycota samples appearing middle of ascomycota branch, or larger classes of taxonomy splitting all over the tree, like saccharomycetes branching from middle of sordariomycetes. Most larger taxon groups are paraphyletic with some elements of polyphyletic elements, especially with sordariomycetes that appear middle of other taxons, while smaller ones are monophyletic. Current taxonomic consensus states that the three phyla should be separate from asco- and basidiomycota, and sordario- and

saccharomycetes are in separate on subphylum level and should not appear as monophyletic. Lack of some fungal phyla can be explained samples been from soil, species been rare or them not having the *EPSPS* gene because they are symbionts (glomeromycota, neocallimastigomycota) or parasites (microsporidia) that take aromatic amino acids from other organisms.

3.3.2. Dollon parsimony

The domain dataset was analyzed with the program Count (Csurös 2010). By highlighting domains (family), each branch indicates a putative gain or loss of the a domain based on Dollon parsimony (figure 9). The phylogenetic tree shows that Shikimate DH has been lost multiple times in different branches of phylogenetic tree (Figure 9), consistent with results from previous section. In addition, many domains have been lost during the evolution of fungi (figure 9 and figure 19). Moreover, some new domains have been acquired recently, and independently, in few species (figure 23) and other are acquired internal lineages and subsequently lost in some branches (figure 9, figure 21-22). Notice that the analysis used in Count does not show when domains have been duplicated, but these cases can be visualized in appendix IV.

The conclusion from this Dollon parsimony analysis is that a 'Full6' multidomain, i.e. a protein sequence with all 6 most abundant domains (table 2), was the original EPSPS sequence in fungi. Moreover, we suggest the hypothesis that sequences with the "Major5" multidomains originated by loss of shikimate DH independently in multiple different branches of evolutionary tree. This has happened both in early evolution (in lower branches) and recently (in higher branches). In some proteins, the shikimate DH (located at the C-term of the protein) is fractured, thus producing non-functional protein. It is possible that this domain was lost in a crossing over event without affecting the shikimate pathway, however we do not know if the fragment is found in any other part of genome. Also, smaller multidomains originated from loss of other shikimate domains, usually in higher branches. Moreover, inclusion of rare domains also occurred in higher branches.

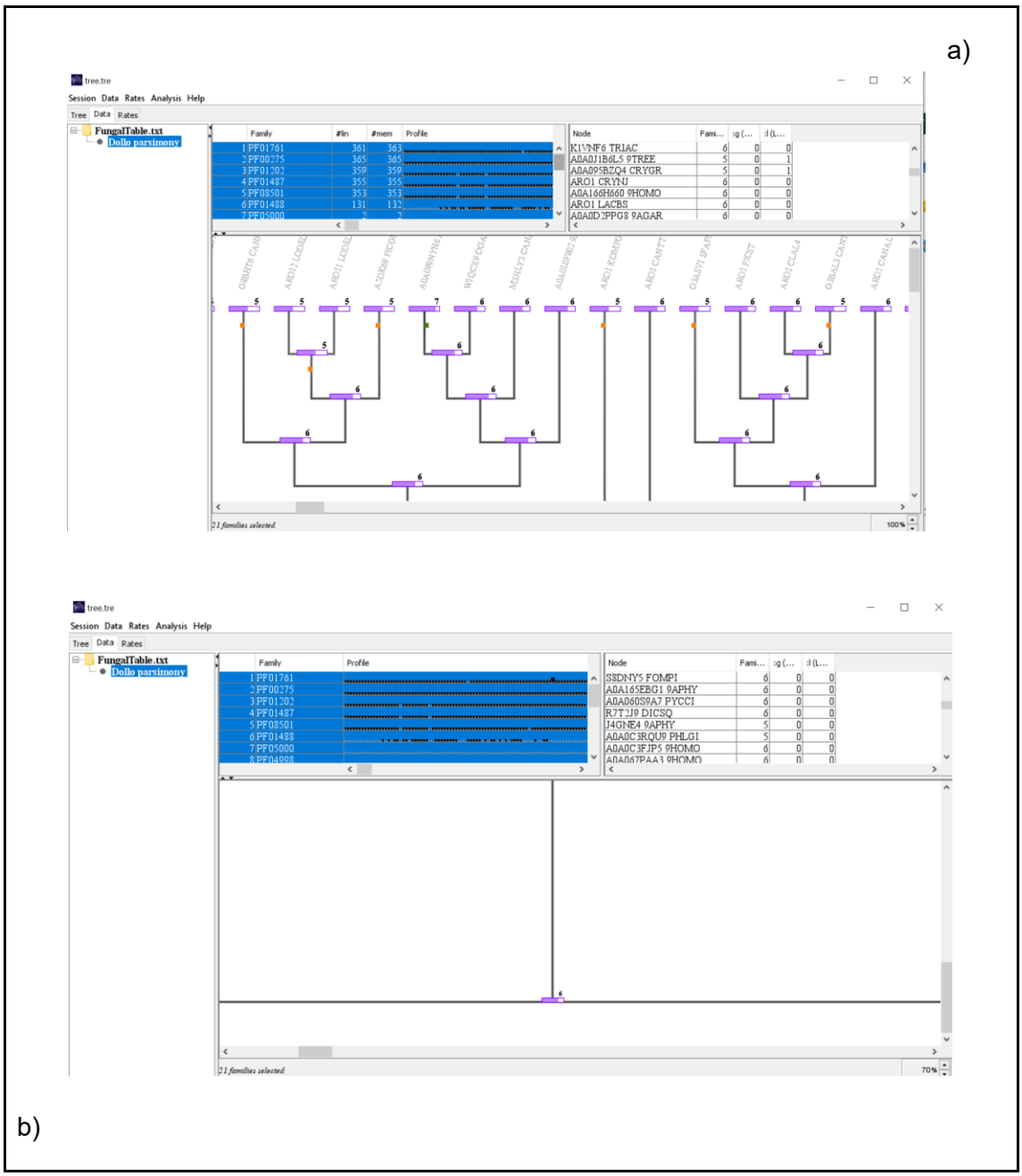

Figure 9. Phylogenetic tree of EPSPS associated multidomain in COUNT with Dollon parsimony. **a)** Picture shows a part of phylogenetic tree. Top left is family (domains) information with the number of terminal ligands in which domains is part of and number of homologs i.e. sum of family size. Profile is visual representation of families. On upper right is node (sample) information with number of families (domains) inside of node and number of lost or gained families leading to the node. On phylogenetic tree orange color on branch indicates loss of domain and green gaining a new one. On leftmost branch loses one domain and the branch from fifth from left gains a new domain. **b)** Lowest part of phylogenetic tree. Full6 is predicted to be original genotype of multidomain. More examples in Appendix IV.

4. CONCLUSIONS

The *EPSPS* gene is found in prokaryotes (archaea and bacteria) and eukaryotes (mostly plants and fungi). The gene encodes for the EPSPS protein that is essential enzyme to

produce aromatic amino acids in the shikimate pathway. In prokaryotes and plants, the *EPSPS* gene is a single domain, but in fungi it is part of multidomain.

The EPSPS multidomain contains multiple domains of shikimate pathway enzymes. It is spread across the fungi, as it is an essential part of protein synthesis. The EPSPS multidomain protein has multiple architectures, from just 2 EPSPS domains to 8 domains. Nevertheless, the most abundant domain architecture (the "Major5") contains DHQ synthase, EPSPS, SKI, DHquinase I and Shikimate DH N).

Based on the visual information from phylogenetic tree and dollon parsimony analysis, the original form of multidomain is the "Full6" architecture, which includes DHQ synthase, EPSPS, SKI, DHquinase I, Shikimate dh N and Shikimate DH. Different branches have lost the shikimate DH domain (independently) to form the "Major5" multidomain structure.

The evolution of the EPSPS protein in fungi shows multiple gains and losses of domains, recent and ancient, as well as some domain duplications.

A scientific article that includes data from this master thesis is publicly available as a preprint at biorxiv (Leino et al. 2020) and submitted to a peer-reviewed journal (appendix VII). Moreover, we submitted an abstract to present at the 1st International Electronic Conference on Genes: Theoretical and Applied Genomics (appendix VIII).

5. FUTURE IMPROVEMENTS

In the future, I would study the products of the EPSPS multidomain to see if each domain produces a separate protein, if they form multidomain protein like AROM complex or if fungi have EPSPS associated domain of shikimate pathway outside of the multidomain (especially in those species that the multidomain architecture includes only 2 or 3 domains). Moreover, I would like to finalize the study of the effect of fungus resistance to the herbicide based on differential domain architectures.

6. ACKNOWLEDGMENTS

I thank Dr. Pere Puigbò for resources, guidance and patience. Also thank Harri Savilahti, Christina Nokkala, Minna Vainio and other lectures of UTU for getting me this far.

I also thank numerous researches who has gathered the materials and shared with the world for free.

And thanks to Mauri Tall and Marjut Lehtonen for producing, rising, and pushing me out of my comfort zone, so I can write these words.

7. REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215(3), pp. 403–410.

Atwood, D. and Paisley-Jones, C. 2017. *Pesticide Industry Sales and Usage — 2008-2012 Market Estimates*. U.S. Environmental Protection Agency.

Balbuena, M.S., Tison, L., Hahn, M.-L., Greggers, U., Menzel, R. and Farina, W.M. 2015. Effects of sublethal doses of glyphosate on honeybee navigation. *The Journal of Experimental Biology* 218(Pt 17), pp. 2799–2805.

Barrera, A., Alastruey-Izquierdo, A., Martín, M.J., Cuesta, I. and Vizcaíno, J.A. 2014. Analysis of the protein domain and domain architecture content in fungi and its application in the search of new antifungal targets. *PLoS Computational Biology* 10(7), p. e1003733.

Barry, G.F., Kishore, G.M., Padgette, S.R. and Stalling, W.C. 1997. Glyphosate-tolerant 5-enolpyruvylshikimate-3-phosphate synthases.

Basu, M.K., Carmel, L., Rogozin, I.B. and Koonin, E.V. 2008. Evolution of protein domain promiscuity in eukaryotes. *Genome Research* 18(3), pp. 449–461.

Bentley, R. 1990. The shikimate pathway--a metabolic tree with many branches. *Critical Reviews in Biochemistry and Molecular Biology* 25(5), pp. 307–384.

Bishop, M.J. and Thompson, E.A. 1986. Maximum likelihood alignment of DNA sequences. *Journal of Molecular Biology* 190(2), pp. 159–165.

Boobies, A.R. 2016. Pesticide residues in food 2016 . *FAO PLANT PRODUCTION AND PROTECTION PAPER* (227), pp. 19–28.

Bundesinstitut für Risikobewertung 2015. The BfR has finalised its draft report for the re-evaluation of glyphosate - BfR [Online]. Available at: https://www.bfr.bund.de/en/the_bfr_has_finalised_its_draft_report_for_the_re_evaluation_of_glyphosate-188632.html [Accessed: 3 June 2020].

Cavallaro, M. 2009. The Seeds Of A Monsanto Short Play [Online]. Available at: https://www.forbes.com/2009/06/29/monsanto-potash-fertilizer-personal-finance-investing-ideas-agrium-mosaic.html#7ee17edb5582 [Accessed: 7 July 2020].

Cerdeira, A.L. and Duke, S.O. 2006. The current status and environmental impacts of glyphosate-resistant crops: a review. *Journal of Environmental Quality* 35(5), pp. 1633–1658.

Clarke, G., Stilling, R.M., Kennedy, P.J., Stanton, C., Cryan, J.F. and Dinan, T.G. 2014. Minireview: Gut microbiota: the neglected endocrine organ. *Molecular Endocrinology* 28(8), pp. 1221–1238.

Clausing, P., Robinson, C. and Burtscher-Schaden, H. 2018. Pesticides and public health: an analysis of the regulatory approach to assessing the carcinogenicity of glyphosate in the European Union. *Journal of Epidemiology and Community Health* 72(8), pp. 668–672.

Corel, E., Méheust, R., Watson, A.K., McInerney, J.O., Lopez, P. and Bapteste, E. 2018. Bipartite network analysis of gene sharings in the microbial world. *Molecular Biology and Evolution* 35(4), pp. 899–913.

Csurös, M. 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* 26(15), pp. 1910–1912.

Duke, S.O. 2018. The history and current status of glyphosate. *Pest Management Science* 74(5), pp. 1027–1034.

Duncan, K., Edwards, R.M. and Coggins, J.R. 1987. The pentafunctional arom enzyme of Saccharomyces

cerevisiae is a mosaic of monofunctional domains. *The Biochemical Journal* 246(2), pp. 375–386.

Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* 14(9), pp. 755–763.

EFSA 2015. EFSA Glyphosate report. , p. 4.

El-Gebali, S., Mistry, J., Bateman, A., et al. 2019. The Pfam protein families database in 2019. *Nucleic Acids Research* 47(D1), pp. D427–D432.

Europian Union, E. 2015. Glyphosate: EFSA updates toxicological profile | European Food Safety [Online]. Available at: http://www.efsa.europa.eu/en/press/news/151112 [Accessed: 20 May 2020].

Farris, J.S. 1970. Methods for computing wagner trees. *Systematic Biology* 19(1), pp. 83–92.

Felsenstein, J. 1978. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. *Systematic Biology* 27(4), pp. 401–410.

Finn, R.D., Tate, J., Mistry, J., et al. 2008. The Pfam protein families database. *Nucleic Acids Research* 36(Database issue), pp. D281-8.

Firdous, S., Iqbal, S., Anwar, S. and Jabeen, H. 2018. Identification and analysis of 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS) gene from glyphosate-resistant Ochrobactrum intermedium Sq20. *Pest Management Science* 74(5), pp. 1184–1196.

Fitch, W.M. 2000. Homology a personal view on some of the problems. *Trends in Genetics* 16(5), pp. 227–231.

Fitch, W.M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic zoology* 20(4), p. 406.

Franz, J.E. 1974. N. PHOSPHONOMETHYLGLYCNE PHYTOTOXCANT COMPOSITIONS.

Funke, T., Han, H., Healy-Fried, M.L., Fischer, M. and Schönbrunn, E. 2006. Molecular basis for the herbicide resistance of Roundup Ready crops. *Proceedings of the National Academy of Sciences of the United States of America* 103(35), pp. 13010–13015.

Galperin, M.Y., Makarova, K.S., Wolf, Y.I. and Koonin, E.V. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research* 43(Database issue), pp. D261-9.

Giesy, J., Dupson S, Solomon, KR 2000. Ecotoxilogical Risk Assement for Roundup Herbicide. *Reviews of Environmental Contamination and Toxicology* 167, pp. 35–120.

Gould, S.J. 1970. Dollo on Dollo's law: irreversibility and the status of evolutionary laws. *Journal of the history of biology* 3, pp. 189–212.

Green, J.M. and Owen, M.D.K. 2011. Herbicide-resistant crops: utilities and limitations for herbicide-resistant weed management. *Journal of Agricultural and Food Chemistry* 59(11), pp. 5819–5829.

Guyton, K.Z., Loomis, D., Grosse, Y., et al. 2015. Carcinogenicity of tetrachlorvinphos, parathion, malathion, diazinon, and glyphosate. *The Lancet Oncology* 16(5), pp. 490–491.

Hasan, N. and Nester, E.W. 1978. Dehydroquinate synthase in Bacillus subtilis. An enzyme associated with chorismate synthase and flavin reductase. *The Journal of Biological Chemistry* 253(14), pp. 4999–5004.

Hawkins, A.R., Moore, J.D. and Adeokun, A.M. 1993. Characterization of the 3-dehydroquinase domain of the pentafunctional AROM protein, and the quinate dehydrogenase from Aspergillus nidulans, and the overproduction of the type II 3-dehydroquinase from neurospora crassa. *The Biochemical Journal* 296 (Pt 2), pp. 451–457.

Herrmann, K.M. 1995. The shikimate pathway: early steps in the biosynthesis of aromatic compounds. *The Plant Cell* 7(7), pp. 907–919.

Herrmann, K.M. and Weaver, L.M. 1999. The shikimate pathway. *Annual review of plant physiology and plant molecular biology* 50, pp. 473–503.

Hibbett, D.S., Binder, M., Bischoff, J.F., et al. 2007. A higher-level phylogenetic classification of the Fungi. *Mycological Research* 111(Pt 5), pp. 509–547.

IARC 2015. IARC Monographs Volume 112: evaluation of five organophosphate insecticides and herbicides. , p. 2.

Iranzo, J., Koonin, E.V., Prangishvili, D. and Krupovic, M. 2016. Bipartite Network Analysis of the Archaeal Virosphere: Evolutionary Connections between Viruses and Capsidless Mobile Elements. *Journal of Virology* 90(24), pp. 11043–11055.

Iranzo, J., Puigbò, P., Lobkovsky, A.E., Wolf, Y.I. and Koonin, E.V. 2016. Inevitability of genetic parasites. *Genome Biology and Evolution* 8(9), pp. 2856–2869.

Jaworski, E.G. 1972. Mode of action of N-phosphonomethylglycine. Inhibition of aromatic amino acid biosynthsis. *Journal of Agricultural and Food Chemistry* 20(6), pp. 1195–1198.

Leino, L., Tall, T., Helander, M., et al. 2020. Classification of the glyphosate target enzyme (5-enolpyruvylshikimate-3-phosphate synthase). *BioRxiv*.

Letunic, I. and Bork, P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23(1), pp. 127–128.

Letunic, I. and Bork, P. 2019. Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research* 47(W1), pp. W256–W259.

Light, S.H., Krishna, S.N., Minasov, G. and Anderson, W.F. 2016. An Unusual Cation-Binding Site and Distinct Domain-Domain Interactions Distinguish Class II Enolpyruvylshikimate-3-phosphate Synthases. *Biochemistry* 55(8), pp. 1239–1245.

Lin, Q., Fan, S., Zhang, Y., et al. 2016. The seahorse genome and the evolution of its specialized morphology. *Nature* 540(7633), pp. 395–399.

Liu, J.-S., Cheng, W.-C., Wang, H.-J., Chen, Y.-C. and Wang, W.-C. 2008. Structure-based inhibitor discovery of Helicobacter pylori dehydroquinate synthase. *Biochemical and Biophysical Research Communications* 373(1), pp. 1–7.

Lumsden, J. and Coggins, J.R. 1977a. The subunit structure of the arom multienzyme complex of Neurospora crassa. A possible pentafunctional polypeptide chain. *The Biochemical Journal* 161(3), pp. 599–607.

Lumsden, J. and Coggins, J.R. 1977b. The Subunit Structure of the arom Multienzyme Complex of Neurospora crassa. *The Biochemical Journal* 161(3), pp. 599–607.

Maeda, H. and Dudareva, N. 2012. The shikimate pathway and aromatic amino Acid biosynthesis in plants. *Annual review of plant biology* 63, pp. 73–105.

Marshall, C.R., Raff, E.C. and Raff, R.A. 1994. Dollo's law and the death and resurrection of genes. *Proceedings of the National Academy of Sciences of the United States of America* 91(25), pp. 12283–12287.

Mesnage, R. and Antoniou, M.N. 2017. Facts and fallacies in the debate on glyphosate toxicity. *Frontiers in public health* 5, p. 316.

Moore, J.D., Coggins, J.R., Virden, R. and Hawkins, A.R. 1994. Efficient independent activity of a monomeric, monofunctional dehydroquinate synthase derived from the N-terminus of the pentafunctional AROM protein of Aspergillus nidulans. *The Biochemical Journal* 301 (Pt 1), pp. 297–304.

Myers, J.P., Antoniou, M.N., Blumberg, B., et al. 2016. Concerns over use of glyphosate-based herbicides and risks associated with exposures: a consensus statement. *Environmental Health: A Global Access Science Source* 15, p. 19.

Naomi, K. 2018. Bayer Closes Monsanto Deal to Cap $63 Billion Transformation - Bloomberg. *Bloomberg*.

Negron, L., Patchett, M.L. and Parker, E.J. 2011. Expression, Purification, and Characterisation of Dehydroquinate Synthase from Pyrococcus furiosus. *Enzyme research* 2011, p. 134893.

Pavlopoulos, G.A., Kontou, P.I., Pavlopoulou, A., Bouyioukos, C., Markou, E. and Bagos, P.G. 2018. Bipartite graphs in systems biology and medicine: a survey of methods and applications. *GigaScience* 7(4), pp. 1–31.

Pleasants, J.M. and Oberhauser, K.S. 2013. Milkweed loss in agricultural fields because of herbicide use:

effect on the monarch butterfly population. *Insect Conservation and Diversity* 6(2), pp. 135–144.

Pollegioni, L., Schonbrunn, E. and Siehl, D. 2011. Molecular basis of glyphosate resistance-different approaches through protein engineering. *The FEBS Journal* 278(16), pp. 2753–2766.

Priestman, M.A., Funke, T., Singh, I.M., Crupper, S.S. and Schönbrunn, E. 2005. 5-Enolpyruvylshikimate-3-phosphate synthase from Staphylococcus aureus is insensitive to glyphosate. *FEBS Letters* 579(3), pp. 728–732.

Richards, T.A., Dacks, J.B., Campbell, S.A., et al. 2006. Evolutionary origins of the eukaryotic shikimate pathway: gene fusions, horizontal gene transfer, and endosymbiotic replacements. *Eukaryotic Cell* 5(9), pp. 1517–1531.

Rogozin, I.B., Wolf, Y.I., Babenko, V.N. and Koonin, E.V. 2006. Dollo parsimony and the reconstruction of genome evolution. In: *Parsimony, phylogeny, and genomics*. Oxford University Press, pp. 190–200.

Sammut, S.J., Finn, R.D. and Bateman, A. 2008. Pfam 10 years on: 10,000 families and still growing. *Briefings in Bioinformatics* 9(3), pp. 210–219.

Schinasi, L. and Leon, M.E. 2014. Non-Hodgkin lymphoma and occupational exposure to agricultural pesticide chemical groups and active ingredients: a systematic review and meta-analysis. *International Journal of Environmental Research and Public Health* 11(4), pp. 4449–4527.

Schönbrunn, E., Eschenburg, S., Shuttleworth, W.A., et al. 2001. Interaction of the herbicide glyphosate with its target enzyme 5-enolpyruvylshikimate 3-phosphate synthase in atomic detail. *Proceedings of the National Academy of Sciences of the United States of America* 98(4), pp. 1376–1380.

Seymour, P., Schrijver, A. and Diestel, R. 2016. Graph Theory. *Oberwolfach Reports* 173.

Shannon, P., Markiel, A., Ozier, O., et al. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* 13(11), pp. 2498–2504.

Silar, P. 2016. Protistes Eucaryotes: Origine, Evolution et Biologie des Microbes Eucaryotes. *HAL Archive*, p. 462.

Sobel, E. and Martinez, H.M. 1986. A multiple sequence alignment program. *Nucleic Acids Research* 14(1), pp. 363–374.

Sonnhammer, E.L., Eddy, S.R. and Durbin, R. 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 28(3), pp. 405–420.

Soumis, N., Reviewers Équiterre, P.D., Ross, K., et al. 2018. *GLYPHOSATE: THE WORLD'S MOST WIDELY USED HERBICIDE 2 Author Canadian Association of Physicians for the Environment*. CAPE.

Starcevic, A., Akthar, S., Dunlap, W.C., et al. 2008. Enzymes of the shikimic acid pathway encoded in the genome of a basal metazoan, Nematostella vectensis, have microbial origins. *Proceedings of the National Academy of Sciences of the United States of America* 105(7), pp. 2533–2537.

Steinrücken, H.C. and Amrhein, N. 1980. The herbicide glyphosate is a potent inhibitor of 5-enolpyruvylshikimic acid-3-phosphate synthase. *Biochemical and Biophysical Research Communications* 94(4), pp. 1207–1212.

Stratonovich, R.L. 1960. Conditional Markov Processes. *Theory of Probability and Its Applications* 5(2), pp. 156–178.

Sutton, K.A., Breen, J., Russo, T.A., Schultz, L.W. and Umland, T.C. 2016. Crystal structure of 5-enolpyruvylshikimate-3-phosphate (EPSP) synthase from the ESKAPE pathogen Acinetobacter baumannii. *Acta crystallographica. Section F, Structural biology communications* 72(Pt 3), pp. 179–187.

Tarazona, J.V., Court-Marques, D., Tiramani, M., et al. 2017. Glyphosate toxicity and carcinogenicity: a review of the scientific basis of the European Union assessment and its differences with IARC. *Archives of Toxicology* 91(8), pp. 2723–2743.

Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* 28(1), pp. 33–36.

Tatusov, R.L., Koonin, E.V. and Lipman, D.J. 1997. A genomic perspective on protein families. *Science*

278(5338), pp. 631–637.

Techopedia 2017. What is a Bipartite Graph? - Definition from Techopedia [Online]. Available at: https://www.techopedia.com/definition/17941/bipartite-graph [Accessed: 16 August 2019].

Tedersoo, L., Sánchez-Ramírez, S., Kõljalg, U., et al. 2018. High-level classification of the Fungi and a tool for evolutionary ecological analyses. *Fungal diversity* 90(1), pp. 135–159.

The European Bioinformation Institute and The European Molecular Biology Laboratory What are protein families? | EMBL-EBI Train online [Online]. Available at: https://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/protein-classification/what-are-protein-families [Accessed: 8 June 2020].

Vandenberg, L.N., Blumberg, B., Antoniou, M.N., et al. 2017. Is it time to reassess current safety standards for glyphosate-based herbicides? *Journal of Epidemiology and Community Health* 71(6), pp. 613–618.

WHO 2019. Glyphosate. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans.* (112), p. 92.

Xu, Q. and Dunbrack, R.L. 2012. Assignment of protein sequences to existing domain and family classification systems: Pfam and the PDB. *Bioinformatics* 28(21), pp. 2763–2772.

Ye, S., Von Delft, F., Brooun, A., Knuth, M.W., Swanson, R.V. and McRee, D.E. 2003. The crystal structure of shikimate dehydrogenase (AroE) reveals a unique NADPH binding mode. *Journal of Bacteriology* 185(14), pp. 4144–4151.

Young, V.R. 1994. Adult amino acid requirements: the case for a major revision in current recommendations. *The Journal of Nutrition* 124(8 Suppl), pp. 1517S-1523S.

8. APPENDIX

I. Aromatic amino acid biosynthesis.

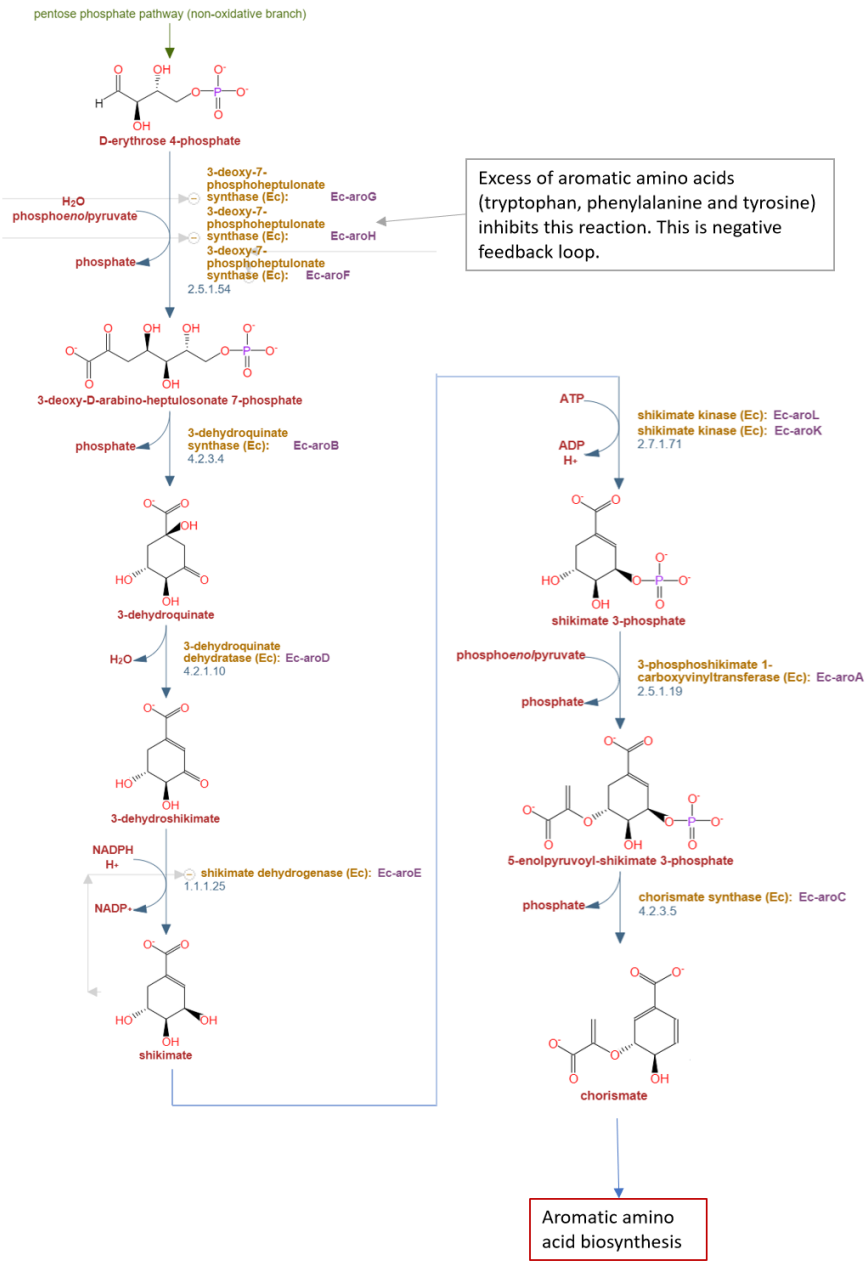

Figure 10. Shikimate Pathway. Part of superpathway of L-phenylalanine, L-tyrosine, and L-tryptophan biosynthesis in detail. Pathway is from *Escherichia coli* K-12 substr. MG1655 species. Structural formula of each substrate and product is shown. Substrates are shown in red text and enzymes in orange text. Purple text is the name of gene that produces the enzyme and blue numbers annotate the reaction. Blue arrows show reaction path. Cray arrows show inhibition (not shown fully due to size); products on biosynthesis suppress their own synthesis: a negative feedback loop. Pathway is cut to figure 10 and figure 11 to fit them to pages. Source https://metacyc.org/ on 14.03.2020.

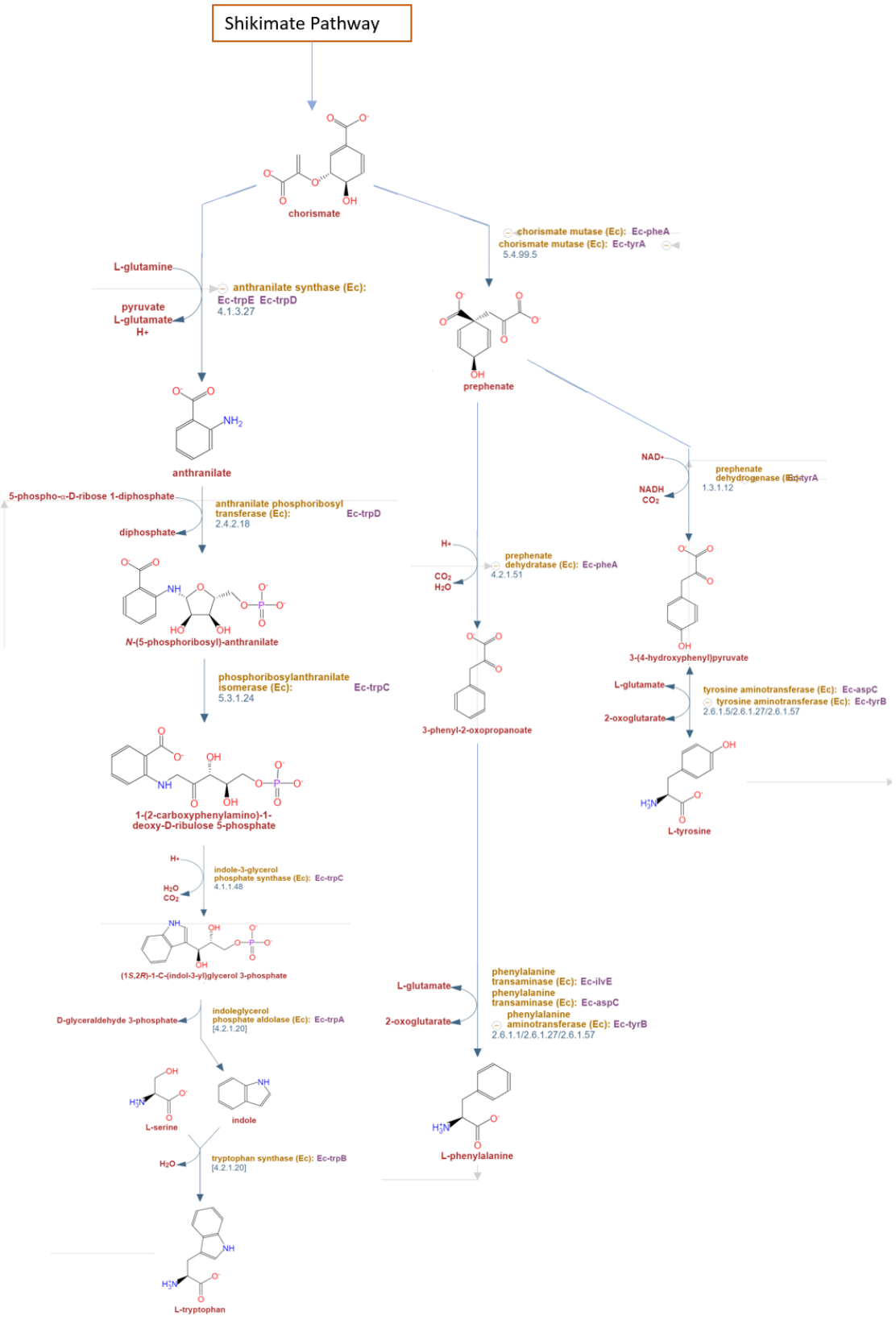

Figure 11. Aromatic Amino Acid Biosynthesis. Part of superpathway of L-phenylalanine, L-tyrosine, and L-tryptophan biosynthesis in detail. Pathway is from *Escherichia coli* K-12 substr. MG1655 species. Structural formula of each substrate and product is shown. Substrates are shown in red text and enzymes in orange text. Purple text is the name of gene that produces the enzyme and blue numbers annotate the

reaction. Blue arrows show reaction path. Cray arrows show inhibition (not shown fully do to size); products on biosynthesis suppress their own synthesis: a negative feedback loop. Pathway is cut to figure 10 and figure 11 to fit them to pages. Source https://metacyc.org/ on 14.03.2020.

II. Multidomain architectures in fungal samples

Table 3. All recorded multidomain architectures and their frequencies. From lowest domain number to highest. The six most important domains have been color coded according to the color scheme from Pfam. The novel domains have been left white. Refer to Table 2 for the function of domains.

1	2	3	4	5	6	7	8	N
EPSPS	EPSPS							2
EPSPS	SKI							4
DHQ S	EPSPS							8
EPSPS	DHQ							1
EPSPS	Methyltransf 11							1
EPSPS	FAD binding 3							1
EPSPS	SDH N	SDH						1
DHQ S	EPSPS	SKI						1
DHQ S	DHQ S	EPSPS	EPSPS					1
DHQ S	EPSPS	SKI	DHQ					2
EPSPS	EPSPS	SKI	SHD N					1
EPSPS	SKI	DHQ	SHD N					5
DHQ S	EPSPS	DHQ	SHD N					1
DHQ S	EPSPS	SKI	DHQ	SDH N				215
EPSPS	SKI	DHQ	SDH N	SDH				3
DHQ S	EPSPS	SKI	SDH N	SDH				1
DHQ S	DHQ S	EPSPS	UCH	SKI				1
DHQ S	DHQ S	EPSPS	SKI	DHQ	SDH N			1
DHQ S	EPSPS	EPSPS	SKI	DHQ	SDH N			2
XRN N	DHQ S	EPSPS	SKI	DHQ	SDH N			2

Glyco hydro 43	DHQ S	EPS PS	SKI	DHQ	SDH N			1
DHQ S	EPSPS	SKI	DHQ	SDH N	SDH			128
DHQ S	EPSPS	SKI	DHQ	SDH N	SDH	Rad52 Rad22		1
RasGAP	CRAL TRIO 2	DHQ S	EPSPS	SKI	DHQ	SDH N		1
GIDA	GIDA assoc	DHQ S	EPSPS	SKI	DHQ	SDH N		1
DUF1977	DHQ S	EPS PS	SKI	DHQ	SDH N	SDH		1
DHQ S	EPSPS	EPS PS	SKI	DHQ	DHQ	SDH N	SDH	1
RNA_pol_Rpb1_1	RNA_pol_Rpb1_2	RNA_pol_Rpb1_3	RNA_pol_Rpb1_4	RNA_pol_Rpb1_5	DHQ S	EPSPS	SKI	2

III. Bipartite graphs

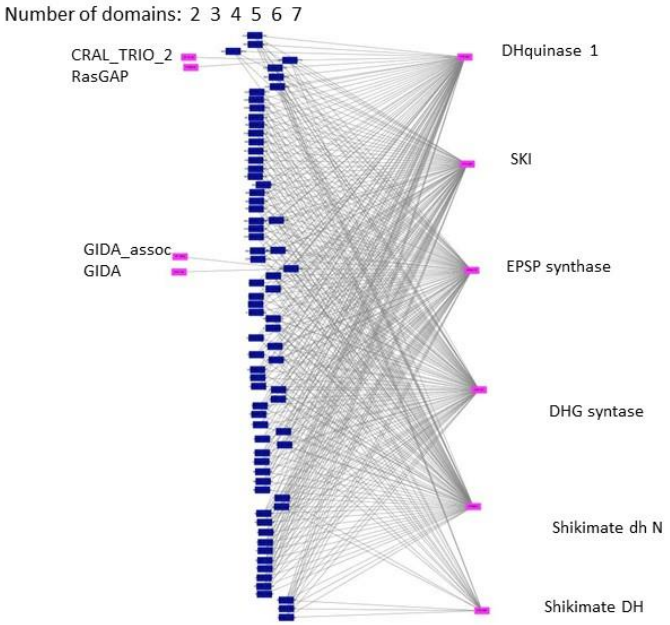

Figure 12: Sordariomycetes. Main six domains are on the left side of samples and the unique domain are on the left side. Domains have been named. Each sample is moved sideways to show the number of domains.

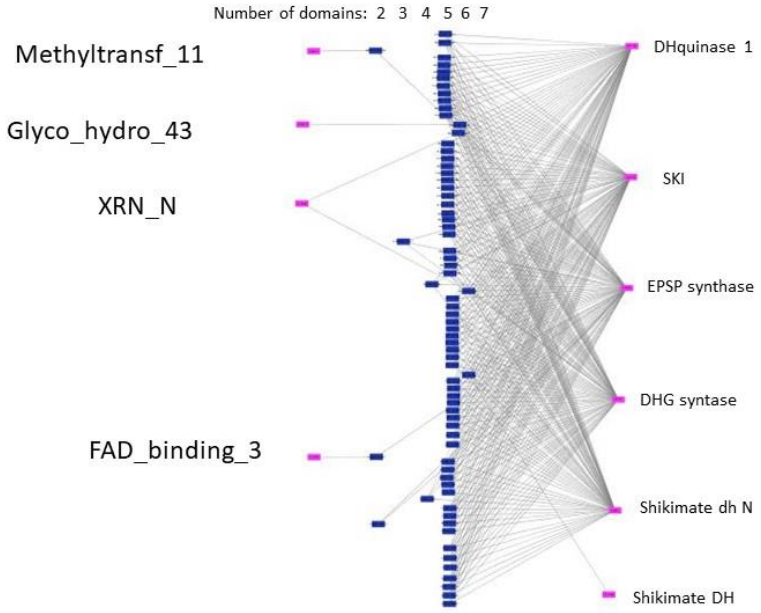

Figure 13. Eurotiomyceres. Main six domains are on the left side of samples and the unique domain are on the left side.Each domain is named. Each sample is moved sideways to show the number of domains.

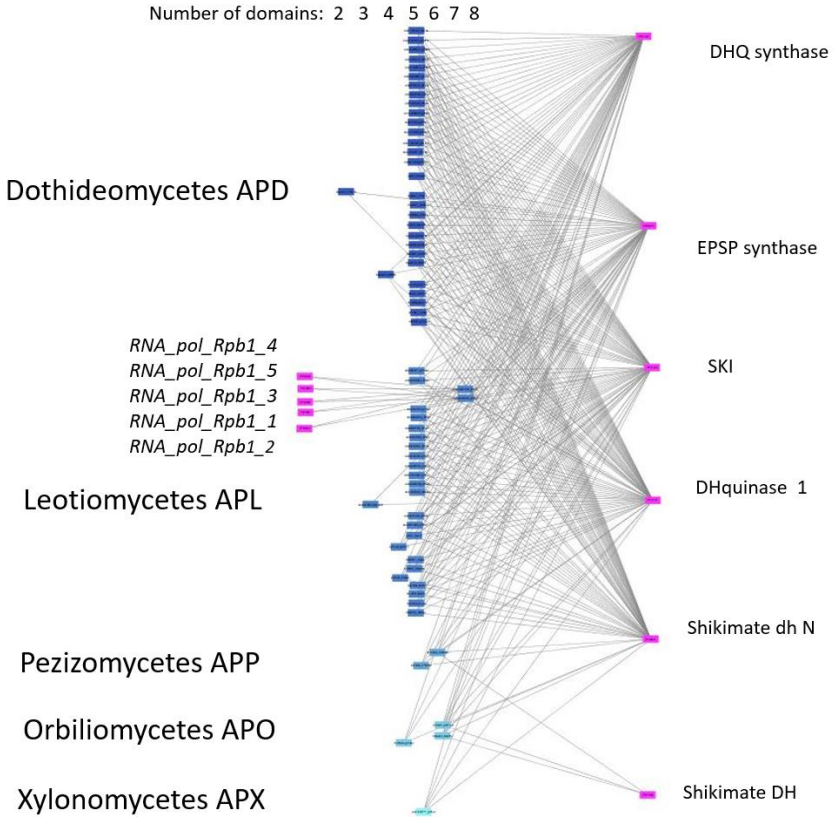

Figure 14. Dothiomycetes, Leotidomycetes, Pezizomycetes, Obriliomycetes and Xylonomycetes classes of samples. Main six domains are on the left side of samples and the unique domain are on the left side. Each sample is moved sideways to show the number of domains.

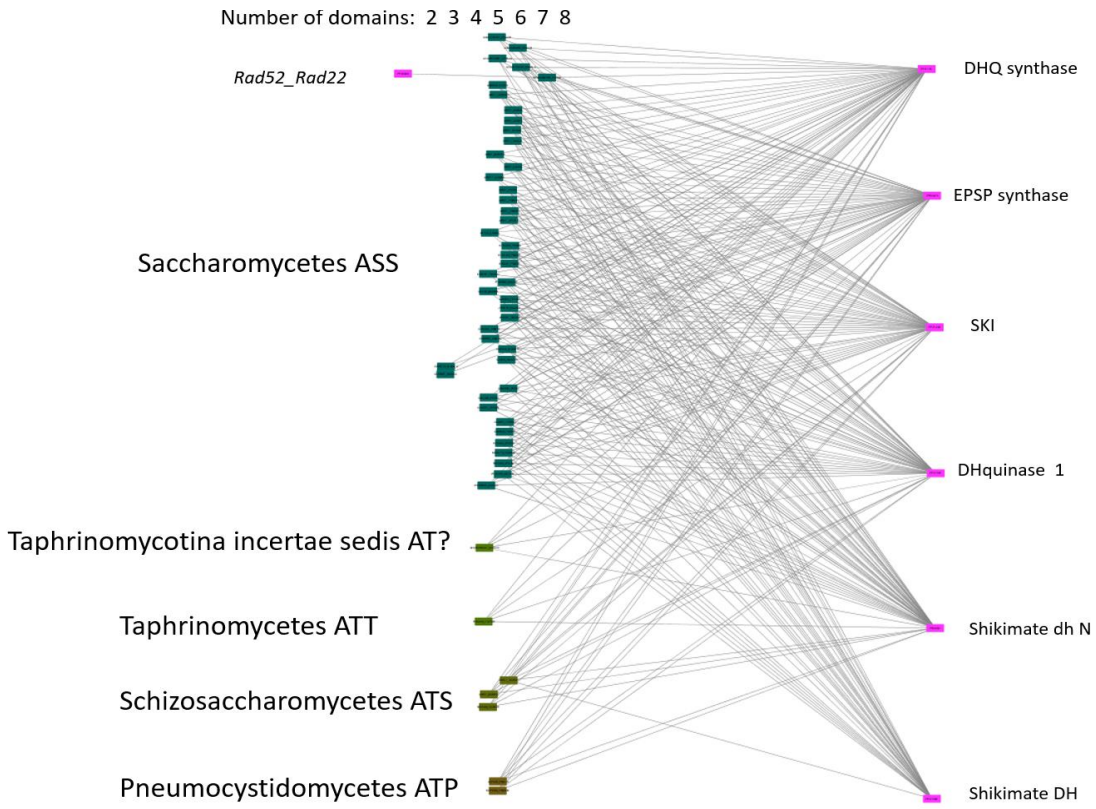

Figure 15. Saccharomycotina, unknown Taphrinomycotina, Taphrinomycetes, Schizosaccharomycete and Pneumocystidomycetes classes of samples. Main six domains are on the left side of samples and the unique domain are on the left side. Each sample is moved sideways to show the number of domains.

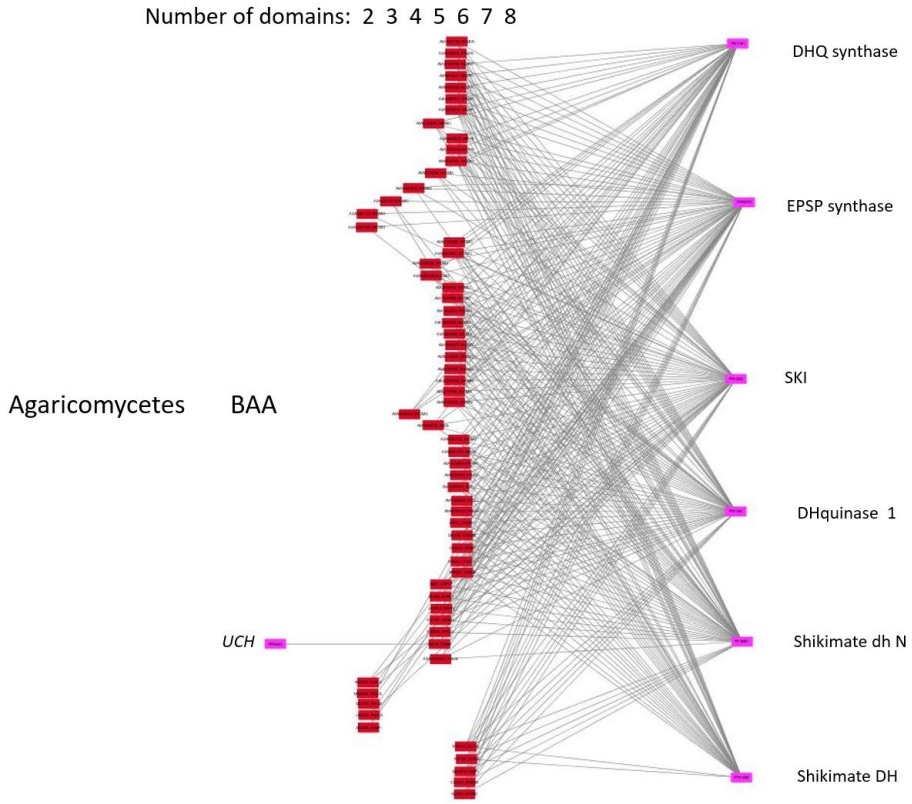

Figure 16. Agaricomycetes. Main six domains are on the left side of samples and the unique domain are on the left side. Each sample is moved sideways to show the number of domains.

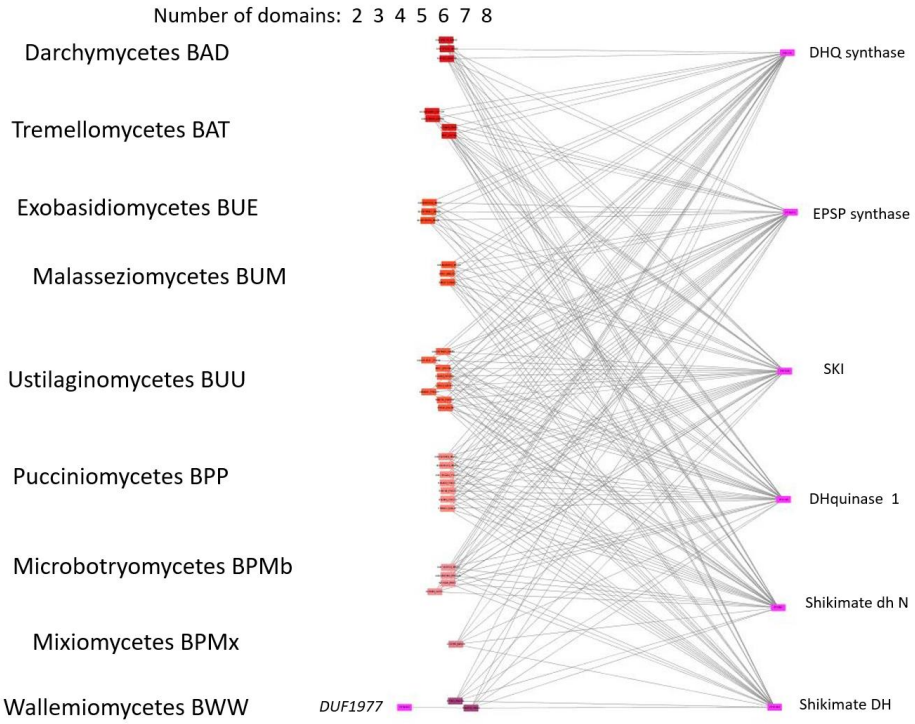

Figure 17. Dacrymycetes Tremellomycetes Exobasidiomycetes Malasseziomycetes and Ustilaginomycetes Pucciniomycetes Microbotryomycetes, Mixiomycetes and Wallemiomycetes. Main six domains are on the left side of samples and the unique domain are on the left side. Each sample is moved sideways to show the number of domains.

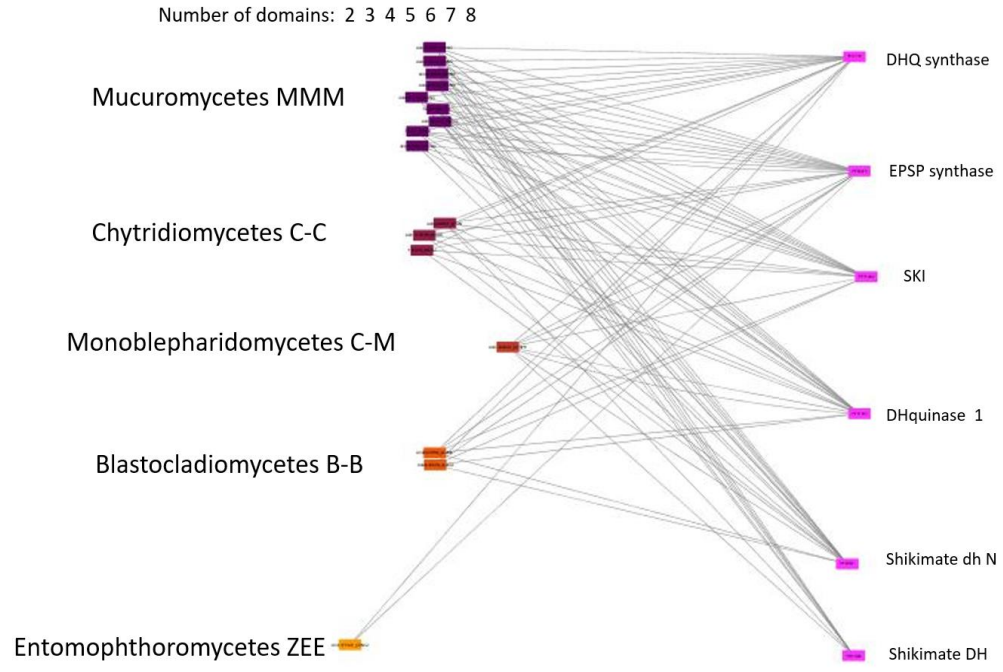

Figure 18. Mucoromycetes, Chytridiomycetes, Monoblepharidomycetes, Blastocladiomycetes and Entomophthoromycetes (ZEE). Main six domains are on the left side of samples and the unique domain are on the left side. Each sample is moved sideways to show the number of domains.

IV. Phylogenetic tree by Count: Dollon Parsimony.

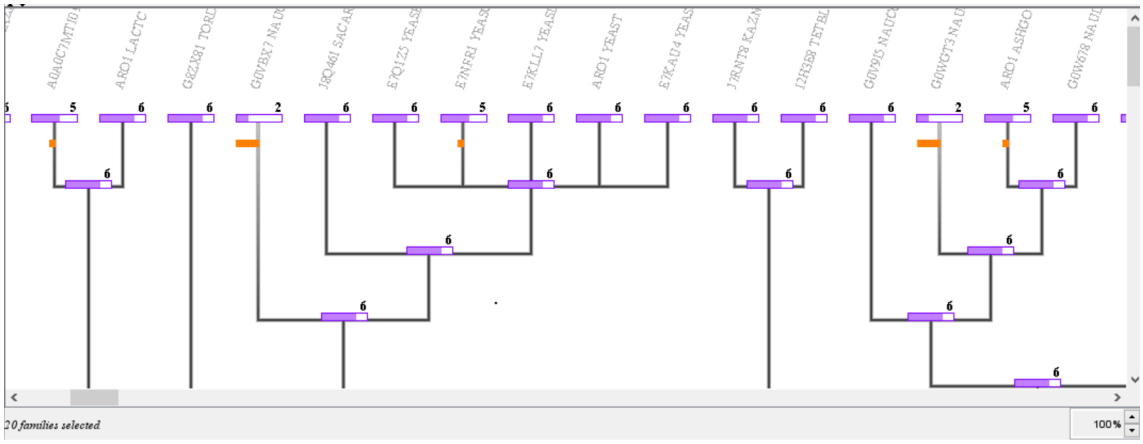

Figure 19. Two events of loss multiple domains. Fourth sample from left (*Naumovozyma castellii*) and third from right (*Naumovozyma dairenensis*) have lost SKI, DHquinase I, Shikimate dh N and Shikimate DH.

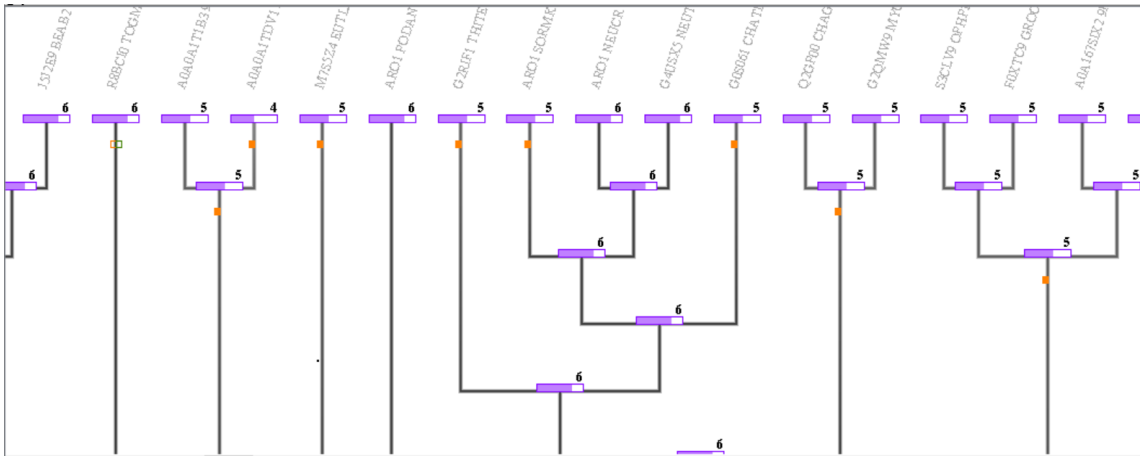

Figure 20. Seeming loss of domain and gaining a new one on second branch from left (*Togninia minima*) but according to Pfam database it has only lost Shikimate DH thus it is Major5 multidomain. Also lose of Shikimate DH to form Major5 multidomains on multiple different branches and lose of Shikimate dh N domain on fourth branch from left (*Torrubiella hemipterigena*).

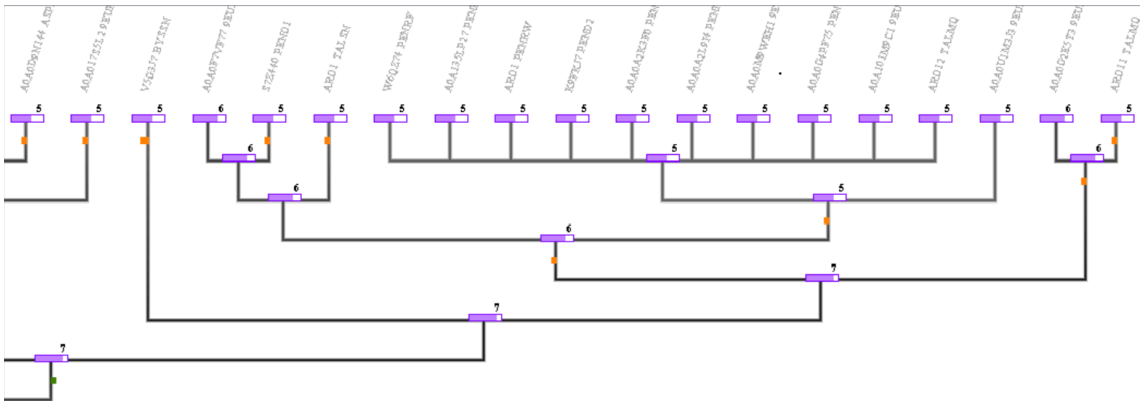

Figure 21. In this branch a new domain (XRN N) is gained in early branch and then lost in most newer branches accept 4th from left.

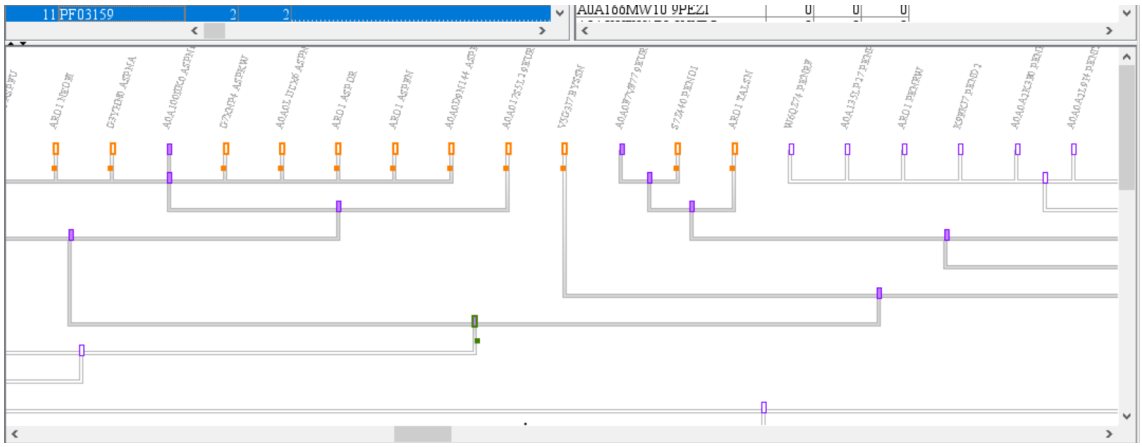

Figure 22. Same area but only the gained domain is highlighted. This domain PF03159 is found in two different branches so it appeared early evolution before it was lost in most branches. It appears at the beginning of multidomain and rest of multidomain is Major5 genotype.

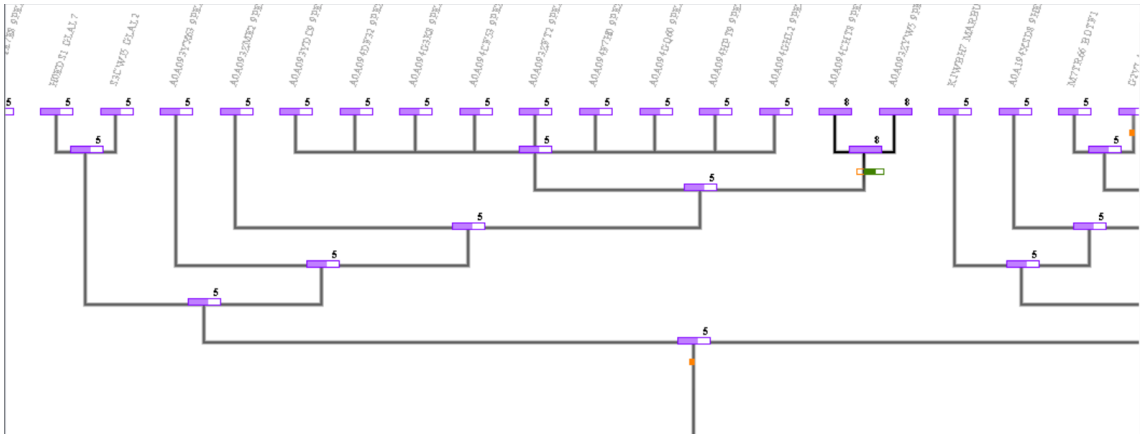

Figure 23. Againing of multiple new domains: RNA_pol_Rpb1_1, RNA_pol_Rpb1_2, RNA_pol_Rpb1_3, RNA_pol_Rpb1_4 and RNA_pol_Rpb1_5, and lose of Shikimate dn N and Shikimate DH domains in the *Pseudogymnoascus* species branch.

V. Phylogenetic tree of the EPSPS domain

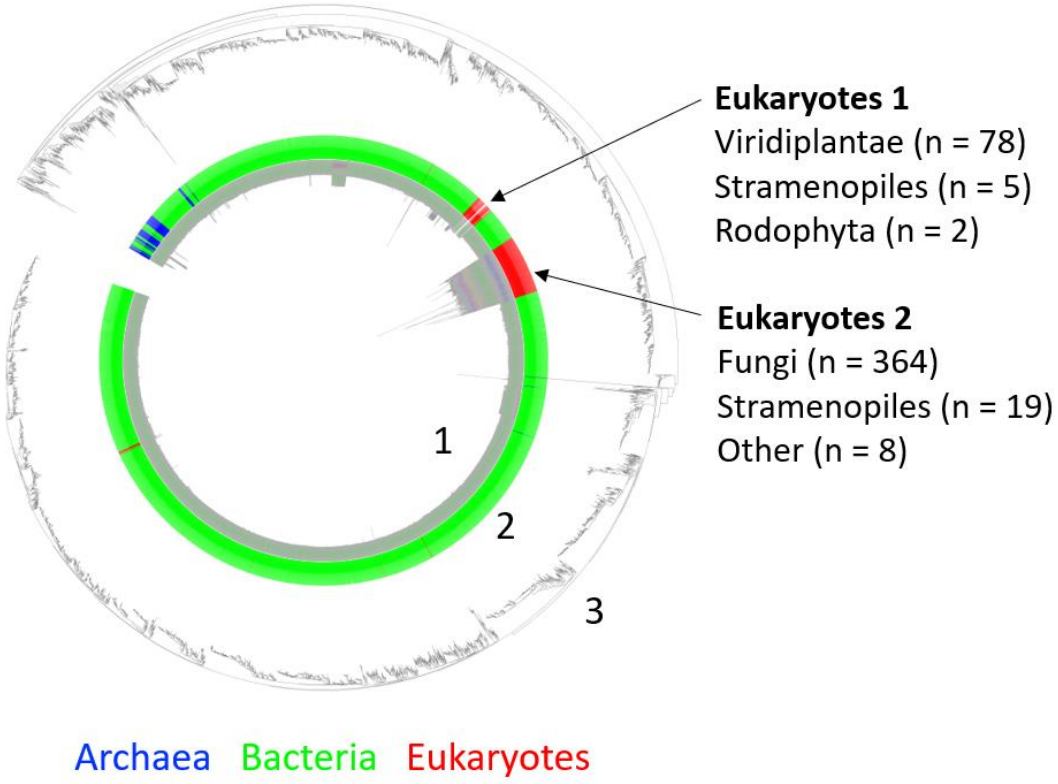

Figure 24. Phylogenetic tree from samples.1) Inner ring is the genes of samples. Multidomains of fungi are seen at Eukaryotes 2. 2) Middle ring is the taxonomy of samples. Bacteria in green, archaea in blue and eukaryotes in red. Eukaryotes are in two groups The first cluster (Eukaryotes 1 contains single-domain sequences corresponding to viridiplantae, stramenopiles and rhodophyta, and the second cluster Eukaryotes 2)contains multidomain proteins (including the EPSPS domain and associated domains) from fungi and a few stramenopiles. 3).

VI. Complete list of EPSP-associated domains in the dataset (n = 46)

Domain	Function	Link	Frequency	Found in species	Freq / Species	Function
PF00275	EPSP_synthase	PF00275	1448	8833	0.16	Produce EPSP Synthase
PF01202	SKI	PF01202	424	7716	0.06	Start of shikimate pathway, phosphorylates shikimate
PF01761	DHQ_synthase	PF01761	420	7429	0.06	Second step of pathway, removes a phosphate from DHAP
PF01487	DHquinase_I	PF01487	416	2073	0.20	3rd step, Hydro-lyase
PF08501	Shikimate_dh_N	PF08501	402	7658	0.05	the substrate binding domain of shikimate dehydrogenase
PF01381	HTH_3	PF01381	218	9670	0.02	Gene expression regulation
PF01488	Shikimate_DH	PF01488	160	6719	0.02	4th step. Dehydrogenesis of shikimate to 5-

						dehydroshikimate
PF02153	PDH	PF02153	127	7136	0.02	Part of shikimate pathway, prephenate dehydrogenases, Tyrosine biosynthesis
PF02224	Cytidylate_kin	PF02224	88	6874	0.01	Kinase of cytidine 5'-monophosphate
PF13193	PF13193	PF13193	17	8031	<0.01	AMP-binding enzyme C-terminal domain for PF00501
PF00501	PF00501	PF00501	17	8544	<0.01	AMP-binding enzyme
PF13560	HTH_31	PF13560	8	5042	<0.01	Helix-turn-helix domain
PF01885	PF01885	PF01885	5	1980	<0.01	RNA 2'-phosphotransferase, Tpt1 catalyses the last step of tRNA splicing in yeast.
PF01817	CM_2	PF01817	3	6634	≪0.01	Chorismate mutase type II: Catalyses the conversion of chorismate to prephenate in the pathway of tyrosine and phenylalanine biosynthesis
PF13419	HAD_2	PF13419	2	9156	≪0.01	Haloacid dehalogenase-like hydrolase
PF07382	HC2	PF07382	2	369	<0.01	bacterial histone H1-like nucleoprotein, DNA condensation
PF05000	RNA_pol_Rpb1_4	PF05000	2	8706	≪0.01	Domain of RNA polymerase
PF04998	RNA_pol_Rpb1_5	PF04998	2	9235	≪0.01	Domain of RNA polymerase
PF04997	RNA_pol_Rpb1_1	PF04997	2	9150	≪0.01	Domain of RNA polymerase
PF04983	RNA_pol_Rpb1_3	PF04983	2	9168	≪0.01	Domain of RNA polymerase
PF03159	XRN_N	PF03159	2	1035	0,19	5'-3' exonuclease for N-terminus
PF00623	RNA_pol_Rpb1_2	PF00623	2	9181	0,02	Domain of RNA polymerase
PF13932	GIDA_assoc	PF13932	1	6635	0,02	Domain at the C-terminus of protein GidA
PF13716	CRAL_TRIO_2	PF13716	1	845	0,12	Protein structural domain that binds small lipophilic molecules
PF13520	AA_permease_2	PF13520	1	6980	0,01	Membrane permeases involved in the transport of amino acids into the cell
PF11706	PF11706	PF11706	1	1426	0,07	C-terminal zinc finger domain.
PF10275	Peptidase_C65	PF10275	1	874	0,11	Highly specific ubiquitin iso-peptidase that removes ubiquitin from proteins
PF09320	DUF1977	PF09320	1	923	0,11	Unknown
PF08241	Methyltransf_11	PF08241	1	8805	0,01	SAM dependent methyltransferases
PF07336	ABATE	PF07336	1	1313	0,08	Stress-induced transcriptional regulator
PF04616	Glyco_hydro_43	PF04616	1	2871	0,03	Arabinanases that hydrolyse the alpha-1,5-linked L-arabinofuranoside backbone of plant cell wall arabinans.
PF04480	DUF559	PF04480	1	2408	0,04	Unknown
PF04098	Rad52_Rad22	PF04098	1	893	0,11	Double-strand break repair protein

PF03819	MazG	PF03819	1	6343	0,02	Maybe pyrophosphohydrolase involved in histidine biosynthesis?
PF03456	uDENN	PF03456	1	788	0,13	Beta domein fot DENN
PF02141	DENN	PF02141	1	937	0,11	A domain which occurs in several proteins involved in Rab- mediated processes or regulation of MAPK signalling pathways
PF01494	FAD_binding_3	PF01494	1	6597	0,02	Monooxygenase, FAD binding in a number of enzymes.
PF01264	Chorismate_synt	PF01264	1	7703	0,01	7th step in pathway, Chorismate synthase
PF01134	GIDA	PF01134	1	6903	0,01	Glucose inhibited division protein A
PF01048	PNP_UDP_1	PF01048	1	8573	0,01	Phosphorylase superfamily
PF01019	G_glu_transpept	PF01019	1	5449	0,02	Transferase (a type of enzyme) that catalyzes the transfer of gamma-glutamyl functional groups.
PF00670	AdoHcyase_NAD	PF00670	1	6317	0,02	S-adenosyl-L-homocysteine hydrolase
PF00616	RasGAP	PF00616	1	890	0,11	All alpha-helical domain that accelerates the GTPase activity of Ras, thereby "switching" it into an "off" position.
PF00443	UCH	PF00443	1	1166	0,09	Ubiquitin carboxyl-terminal hydrolase
PF00132	Hexapep	PF00132	1	9076	0,01	Bacterial transferase hexapeptide
PF00011	HSP20	PF00011	1	7623	0,01	Small heat shock proteins

VII. Article under review in a peer-reviewed journal and publicly available as a preprint at BioRxiv (Leino et al. 2020).

Classification of the glyphosate target enzyme (5-enolpyruvylshikimate-3-phosphate synthase)

Lyydia Leino [1,&], Tuomas Tall [1,&], Marjo Helander [1], Irma Saloniemi [1], Kari Saikkonen [2], Suvi Ruuskanen [1], Pere Puigbò [1, 3, *]

[1] *Department of Biology, University of Turku, Turku, Finland*

[2] *Biodiversity Unit, University of Turku, Finland*

[3] *Nutrition and Health Unit, Eurecat Technology Centre of Catalonia, Reus, Catalonia, Spain*

[&] *Equal contribution of the authors*

[*] *Corresponding author: pepuav@utu.fi*

Keywords: shikimate pathway, epsps enzyme, glyphosate, herbicide, resistance, microbiome, bioinformatics resource, biomarkers

ABSTRACT

Glyphosate is the most common broad-spectrum herbicide. It targets the key enzyme of the shikimate pathway, 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS), which synthesizes three essential aromatic amino acids (phenylalanine, tyrosine and tryptophan) in plants. Because the shikimate pathway is also found in many prokaryotes and fungi, the widespread use of glyphosate may have unsuspected impacts on the diversity and composition of microbial communities, including the human gut microbiome. Here, we introduce the first bioinformatics method to assess the potential sensitivity of organisms to glyphosate based on the type of EPSPS enzyme. We have precomputed a dataset of EPSPS sequences from thousands of species that will be an invaluable resource to advancing the research field. This novel methodology can classify sequences from >90% of eukaryotes and >80% of prokaryotes. A conservative estimate from our results shows that 54% of species in the core human gut microbiome are sensitive to glyphosate.

VIII. Abstract submitted to present at the 1st *International Electronic Conference on Genes: Theoretical and Applied Genomics* (Tall T and Puigbò P, 2020).

The glyphosate target enzyme 5-enolpyruvylshikimate 3-phosphate synthase (EPSPS) contains several EPSPS-associated domains in fungi

Tuomas Tall [1,*], Pere Puigbò [1,2]

[1] Department of Biology, University of Turku (Turku, Finland)
[2] Nutrition and Health Unit, Eurecat Technology Centre of Catalonia (Reus, Catalonia)

[*] Contact e-mail: teltal@utu.fi

Keywords: shikimate pathway, herbicide, glyphosate, domain architecture, fungi

The 5-enolpyruvylshikimate 3-phosphate synthase (EPSPS) is the central enzyme of the shikimate pathway to synthesize three aromatic amino acids in fungi, plants and prokaryotes. Glyphosate is a multi-spectrum herbicide largely utilized to control weeds, which targets the EPSPS enzyme and inhibits the production of these essential amino acids. In most plants and prokaryotes, the EPSPS protein is constituted by a single domain, whereas in fungi contains the EPSPS and several EPSPS-associated domains. Here, we perform a comprehensive analysis of 391 EPSPS proteins of fungi gathered from the Pfam database. We analyze our dataset with a bipartite graph (Cytoscape) and dollon parsimony (Count) to determine the distribution and the evolution of the 22 EPSPS-associated domains in fungi. The EPSPS-associated domains can be classified into four partially overlapping groups: shikimate pathway, other enzymes, gene expression and structural proteins. The most frequent EPSPS-associated domains are shikimate kinase, 3-dehydroquinate synthase, 3-dehydroquinate dehydratase, shikimate dehydrogenase substrate binding domain and shikimate DH. These domains are present in 56% of the proteins analyzed and 34% of proteins contain shikimate DH at the end of sequence. The most common domain architecture of the EPSPS enzyme in fungi contains 5-6 domains. A parsimony analysis suggests that a 6-domain protein is the ancestral form of the EPSPS in fungi and that alternative architectures are due to domain losses (also some gains) and duplications. The results of this study will be useful to determine the impact of glyphosate in fungi and to quantify its putative differential effects on alternative domain architectures.