SYSTEMS BIOLOGY APPROACHES IN THE INVESTIGATION OF ARTICULATION POINTS IN KEGG METABOLIC PATHWAYS

**Running title: INVESTIGATION OF ARTICULATION POINTS**

Igor Brandão1, Alice Câmara2,Diego Arthur1, Leonardo Campos1, Clovis Reis1, Rodrigo JS Dalmolin1\*

1 Bioinformatics multidisciplinary environment, Federal University of Rio Grande do Norte, Natal, Rio Grande do Norte, Brasil

2 Biophysics and Pharmacology Department, Federal University of Rio Grande do Norte, Natal, Rio Grande do Norte, Brasil

\* Corresponding author

E-mail: rodrigo.dalmolin@imd.ufrn.br

# ABSTRACT

The study of proteins essentiality through laboratory methods is expensive, time-consuming and not scalable for large amounts of proteins. Besides, it is relevant to evaluate the essentiality of several proteins of a metabolic pathway as a whole. The metabolic pathways can be analyzed as graphs, which provide several tools to study the topological features such as the articulation points. Nowadays, research in bioinformatics studies the essentiality of proteins based on betweenness and degree metrics, however, graph theory suggests that articulation points could be essential nodes in a network. It remains to be determined whether these articulation points are essential in metabolic pathways and their topological impact on the network. Using network analysis via metrics and biologic curation, we aim to verify if bottlenecks are proteins with the highest frequencies and located in the center of KEGG metabolic pathways. For this purpose, we identified the articulation points in different networks, evaluate the impact of each articulation point, calculate their frequency and compare them with occurrences of non-articulation points. We consulted KEGG pathways available as KGML (KEGG XML) files. After, the data was transformed into a graph object. Two centrality parameters including articulation points and degree are determined and the essential proteins based on these parameters are classified. Approximately 20% of the proteins are articulation points. The articulation points with high-frequency which are located in central regions of the network were considered the most important (3.75%). In addition, the highest concentration of articulation points occurred in the frequency range of 80-90%. A pattern of non-randomness of articulation points was identified in the protein groups that have a frequency of at least 74.5%. Finally, steroid biosynthesis is the metabolic pathway with the highest number of articulation points with frequency higher than 80%. Besides, oxidoreductase is the articulation point class present in the highest number of metabolic pathways. Overall, the findings suggest that bottlenecks are articulation points with highest frequencies and located in the center of the network. It remains to perform a deep analysis on the articulation points biological roles.

**Key-words:** Articulation point, KEGG, Metabolic pathway, Biological network, Systems biology.

# INTRODUCTION

Regulatory and interaction networks are important representations of protein networks and are characterized by having directional and non-directional edges, respectively [1]. The study and application of graph theory in the context of biological networks has gained notoriety since it is necessary to systematically understand how molecules interact with each other and how their functions are determined within the complex cell machinery, alone or together with other cells [2]. Metabolic pathways and cycles are reaction chains where chemical products become the substrate for the next step [3]. Besides, structural and functional analysis of genome-based large-scale metabolic networks is important for understanding the design principles and regulation of metabolism at a system level. The metabolic network is organized into many small and highly connected modules that combine hierarchically into larger, less cohesive units [4]. Therefore, a rational reduction of the metabolic network to its core structure and a deeper understanding of its functional modules are important [5].

The main characteristic of metabolic pathways is that reactions are connected by their intermediates. The products of a reaction are the reactants of subsequent reactions [6]. In general, the metabolic pathways can be classified as flow networks, where a specific variable such as mass or energy flow may be conserved at each node. Metabolic networks have unique properties such as the conservation constraints, which have to be satisfied at each node. Another property is that the metabolic networks are represented with nodes as metabolites and the links are reactions that are catalyzed by specific gene products [7].

In terms of graphs, there are multiple ways to build a network from a metabolic model [8]. Network analysis suggests that biological networks have two important structural properties. First, most of these networks, including metabolic networks are scale-free and possess a ‘‘small-world’’ property, that is characterized by a short average path length [1, 5, 7]. However, it is important to note the existence of driving forces that can influence a novel node attachment in a biological network, such as exon shuffling, retroposition, mobile elements, horizontal gene transfer, gene duplication and the fact that a new node connection reflects its origin with the nature of the node properties [9]. Second, scale-free networks are suggested to have high error tolerance (against random failure) and low attack tolerance (vulnerability to the failure of the highly connected nodes) [7]. The topological features of networks can be measured by observing specific characteristics from each metabolic pathway such as betweenness, degree, articulation points (AP), and bridges. Betweenness centrality measures the total number of non-redundant shortest paths going through a certain node or edge [10]. The degree tells how many links the node has to other nodes [1]. A node is an AP if its removal disconnects the network or increases the number of connected components of the network [11, 12]. In addition, an edge can be classified as a bridge if its removal disconnects the graph or increases the number of connected components otherwise [13].

A great source of curated functional information is the KEGG database (Kyoto Encyclopedia of Genes and Genomes). It holds a knowledge base on metabolic pathway maps of molecular interaction and orthology relationships between genes/gene products [14]. In KEGG, nodes depict enzymes and edges represent the reactions that transform one metabolite into another [15-17]. Although KEGG diagrams are informative and easy to understand, a network topological study about the metabolic pathway can be important to understand specific network characteristics. In this sense, the representation via graphs could stand out because it enables the study of several characteristics such as clusters, articulations points, bridges, an arrangement of proteins in the network, number of protein connections, among other factors. There are many ways to provide visualizations of metabolic pathways.

To date, no research systematically evaluated the topology of all metabolic maps as a whole to study the APs. The study of APs is relevant because by identifying these points it is possible to detect the most vulnerable sites in a metabolic pathway. Besides, the APs can be measured by the number of vertices that get disconnected from the main (largest) surviving connected component after the removal of the AP [18].

In this work, we aim to verify if bottlenecks are proteins with the highest frequencies and located in the center of KEGG metabolic pathways. For this, we analyse which proteins disconnect the network into different components after its remotion. Also, we verify the occurrence of each protein in the set of KEGG metabolic pathways for different organisms available in the KEGG database. Furthermore, we perform the APs frequency evaluation to determine in which frequency range they aren't randomly distributed. Since we detect different profiles for APs proteins in terms of frequency, we separate them into two different groups, the APs with frequencies greater/equal to 80% and lower to 30%. We performed a functional analysis with these AP. Finally, to provide a user-friendly interaction with the assessed networks, we develop an interactive visualization enabling access to each protein topological features and the simulation of creation and remotion of nodes.

# MATERIALS AND METHODS

## Extraction and processing of metabolic pathway data

Initially, the list of all species available in KEGG was generated along with the list of metabolic pathways for each species, and they were the reference for the selection of metabolic pathways used in the study. Each pathway has an associated KGML (KEGG XML) file that contains the enzymes and reactions data necessary to generate the graph object. It is important to note that in some KEGG metabolic pathways there are different nodes with the same enzyme commission number (EC). Since it wasn't possible to identify these nodes uniquely via the EC or KEGG identifier, we needed to combine the following node attributes to differentiate them from each other: EC, reaction code and KEGG diagram (x, y) coordinates.

We used the R/Bioconductor KEGGREST package [19] to load the reference pathway of interest from the REST API provided by KEGG. We used 153 metabolic pathways of all species available on KEGG (totalizing more than 600.000 datasets). KEGG data was parsed using a function adapted from the KEGGREST package. After that, compounds and interconnections between metabolic pathways were removed to study each isolated pathway. We used the R/Bioconductor iGraph package [20] to extract several features from the graph like communities, betweenness centrality, closeness, and clustering. Proteins were classified into two groups according to their topological role in a given metabolic pathway: AP (articulation point), and “Other” (non-articulation point).

## APs detection

APs detection relies on the depth search algorithms (DFS) for graphs. DFS consists of a technique to visit each node in a given graph until its deepest level. Furthermore, each node reached is placed on a stack and a record is kept of the lowest node on the stack to which it is connected by a path of unstacked nodes. Besides, when a new node cannot be reached from the top of the stack, the top node is deleted and the search is continued from the next node on the stack. If the top node of the stack does not connect to a node lower than the second node on the stack, then this second point is an AP of the graph. Finally, all edges examined during the search are placed on another stack. Therefore, when an AP is found, the edges of the corresponding biconnected component may be retrieved and placed in an output array [21].

## Protein frequency evaluation

The frequency evaluation of each protein for each metabolic pathway consists of a merge of all species dataset and protein aggregation in the same pathway with equal attributes to generate the total number of occurrences for a given protein. Besides, it is accounted for the total number of evaluated species as a reference to obtain the protein frequency. The protein total frequency is its occurrence divided by the total evaluated species for a given pathway.

Subsequently, it's generated a dictionary to match the KGML file node with a corresponding EC number, because in a given metabolic pathway it's possible to have two or more nodes with the same EC number. This dictionary contains the attributes pathway code, protein EC number, the reaction code, protein x-axis, and y-axis parsed from the KGML file. After matching the protein attributes with the dictionary, it receives the dictionary ID as a new attribute to be considered a unique node.

## Calculation of AP impacts

The impact coefficient was calculated deleting the APs that affect the network structure and the number of nodes separated from the core component of the network was counted, in other words, the smaller network component. A higher number of disconnected nodes represents a higher impact coefficient for an AP and vice-versa. However, networks with a bigger number of nodes probably would generate APs with higher impacts when compared with another network containing fewer nodes. To solve this issue, a normalization was applied to evaluate the impact coefficient according to the network size. The impact coefficient was normalized on a scale from 0 to 1, 1 represents the highest impact calculated for a given network and 0 represents the lowest impact. The remaining impacts were allocated into intermediate values ​​of the scale.

## Statistical analysis

The datasets processed in the R language containing the total frequency of all nodes were merged and nodes with a zero occurrence were removed to avoid possible bias. Also, the hypergeometric test was performed to verify the distribution and randomness of the APs found in the networks using the *phyper* built-in R function with *lower.tail* parameter set as false. For this purpose, the protein frequency was sorted in descending order to verify if the AP is more concentrated in protein groups with higher occurrences. We performed hypergeometric test N times, with N equal to the number of proteins in the dataset. Each hypergeometric tests the current and the previous proteins in the list in a cumulative way. We normalized the protein frequency according to the total species containing a certain metabolic pathway since the studied pathways showed different numbers of organisms. Besides, we normalized the proteins according to the maximum and minimum frequencies of proteins related to a given metabolic pathway. The highest normalized value was considered 1, the lowest value was considered 0 and others values were within this scale from 0 to 1. P-value < 0.01 was considered significant, showing that the AP proteins were not distributed randomly.

## Functional analysis

To evaluate the main metabolic pathways related to the APs, we performed a study in the KEGG pathway database to understand the APs role in the evaluated metabolic pathways and its functionality at the global level. EC numbers regarding the APs annotated in the database were used to identify the metabolic pathways with a higher number of APs and the APs that are presented in a higher number of pathways.

## Metabolic pathways visualization

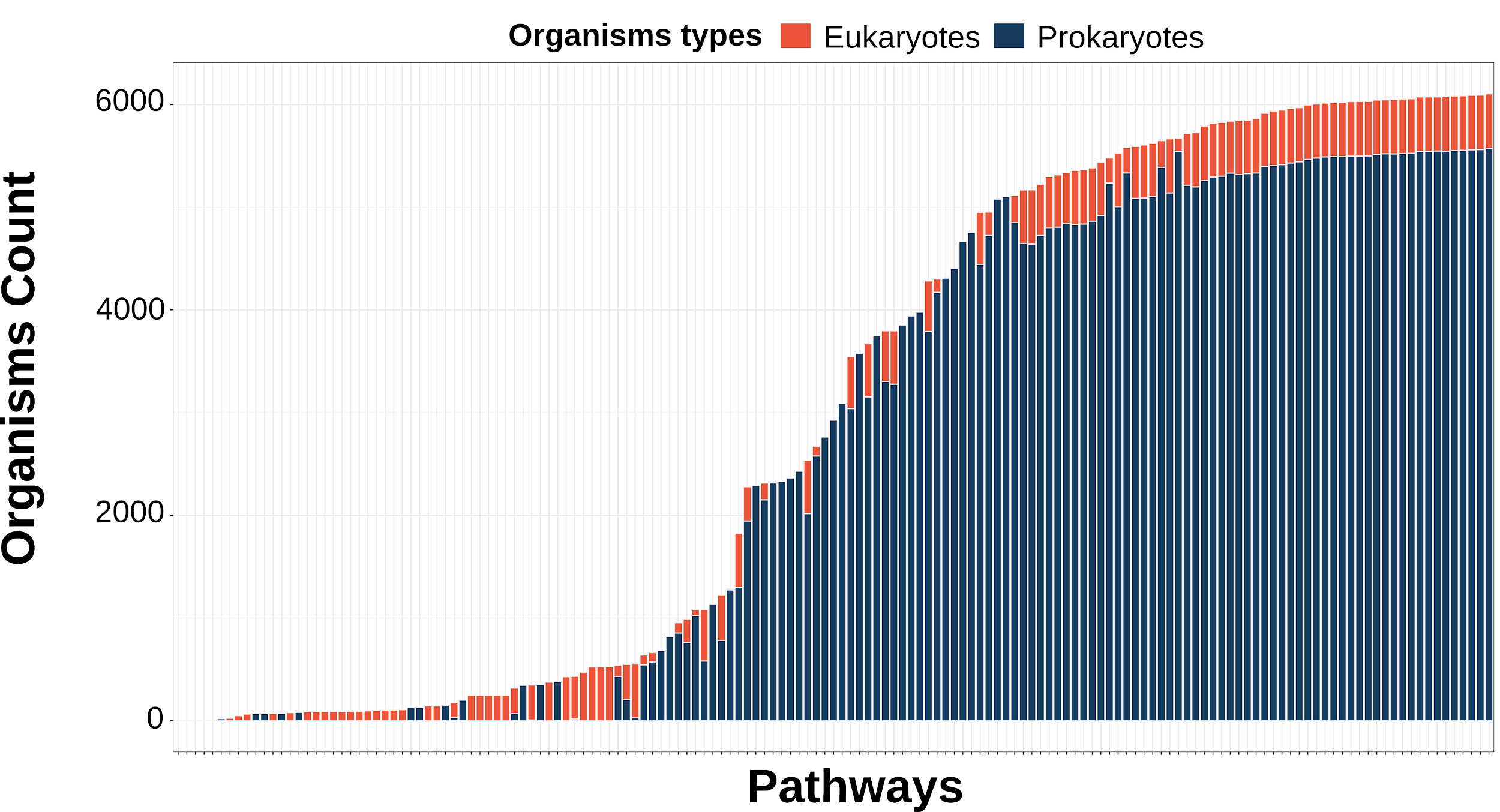
Since a metabolic pathway can be considered a mass flow, we represent it by directed graphs where nodes are the enzymes and edges are the chemical reactions. Also, we provided a graphical visualization of the pathways displaying the main characteristics from its nodes such as AP status (border in blue), betweenness centrality (background color), frequency (node size), protein classification (label), and bridge edges (red arrowed line) using visNetwork R package [22].

# 

# RESULTS

## Number of species in metabolic pathways

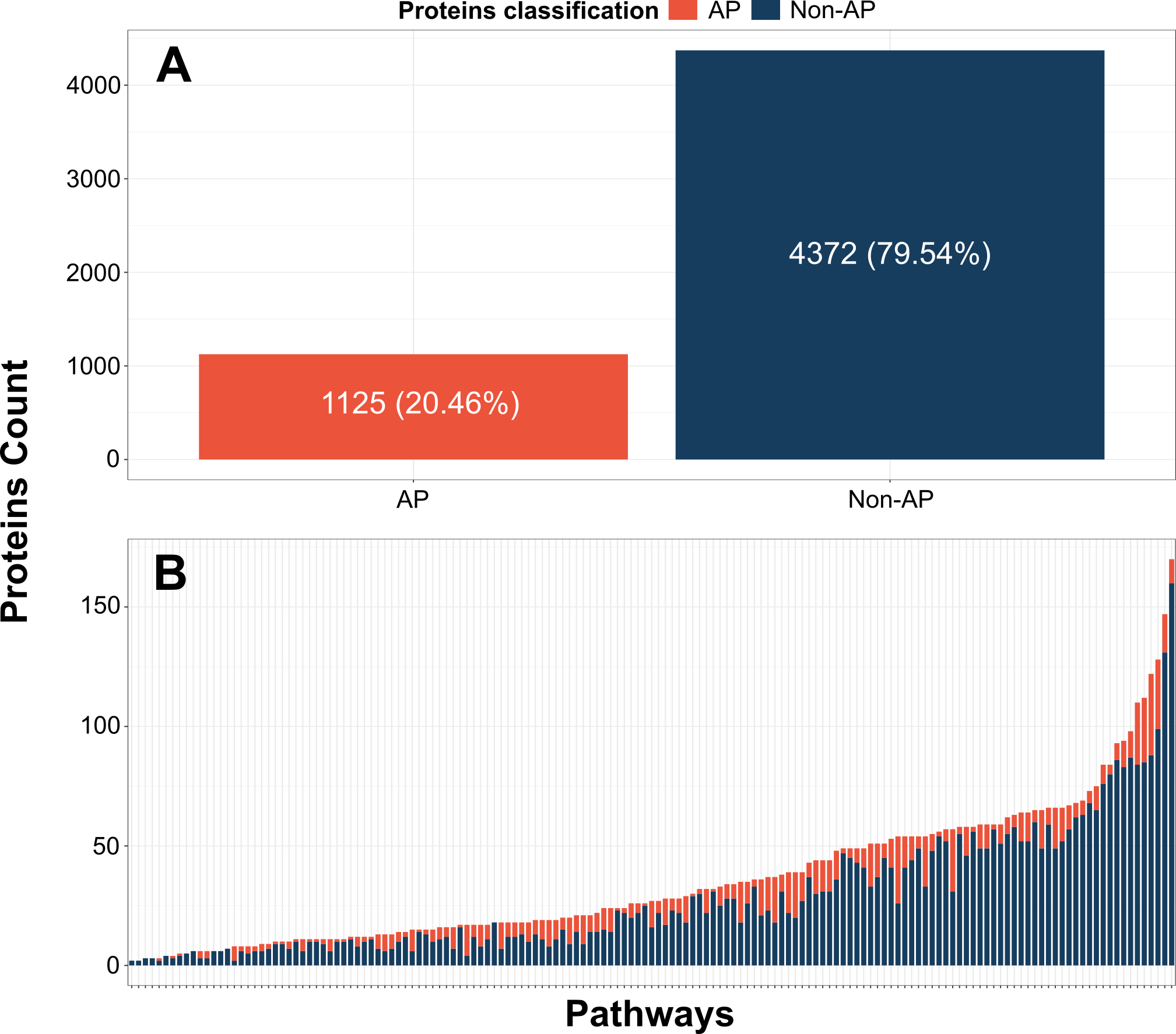
Initially, a script was developed in R language using the KEGGREST package in order to load the list of 6221 organisms with their respective taxonomic information. Through the taxonomy, the organisms were separated according to the kingdoms. Approximately 32% of metabolic pathways are related to less than one thousand species (Figure 1). Additionally, approximately 31% of metabolic pathways are related to at least 5000 species.



**Figure 1:** Distribution of species quantity by metabolic pathways available in KEGG. X-axis: metabolic pathway. Y-axis: organisms count. The orange bars represent pathways related to eukaryotes. Eukaryotes represent 8.6% (534) of total species (6221). Blue bars represent pathways related to prokaryotes. Prokaryotes represent 91.4% (5687) of total species.

## Protein classification

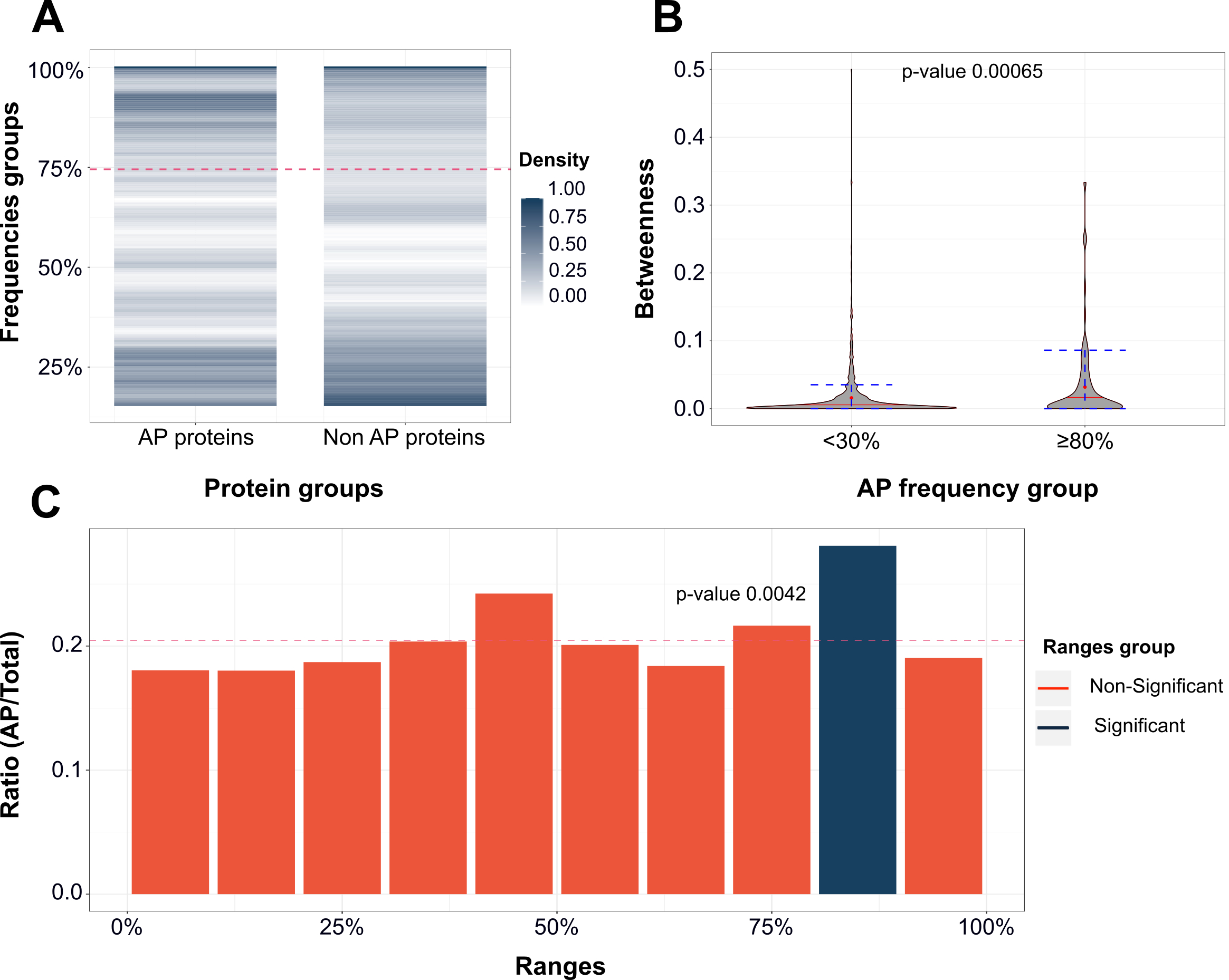
The nodes studied (n=5497) were divided into 2 groups (Figure 2) according to their classification of AP or Non-AP. The APs nodes (n=1125) represented 20.46% of the total evaluated data, while the Non-APs nodes (n=4372) corresponded to 79.54%.



**Figure 2: (A)** Proteins classifications. AP (20.46%); Non-AP (79.54%). Total number of studied nodes: 5497. **(B)** The x-axis represents each pathway and the y-axis depicts the number of nodes per pathway. Protein count per pathway varies from 2 to 170 nodes.

## Articulation points frequency evaluation

Through the hypergeometric analysis, a pattern of non-randomness was identified in the proteins that have a frequency of at least 74.5%, representing the cut-off point of the test. The cut-off point corresponds to the minimum frequency required for a protein to have a non-random distribution. Within this filter, 1047 more frequent proteins were selected, which presented a higher frequency than the cut-off point, of which 236 are APs (22.54%). Regarding the total of studied proteins, 1125 APs (20.46%) were evaluated in which a higher AP density was observed above the cut-off point when compared to non-AP proteins (Figure 3A). Besides, at frequencies groups greater/equal than 80% and lower than 30%, we found the highest concentration of AP (Figure 3A). To differentiate the profile of the APs in the groups of high frequency (>= 80%) and low frequency (<30%), we studied the betweenness of the APs as an indicator of centrality in the assessed metabolic pathways. In this sense, we observed that the APs in the group greater/equal than 80% presented betweenness higher than the group lower than 30% (Figure 3B), suggesting that these APs are located at more strategic points of the metabolic pathways (Supporting material S1). Additionally, we evaluated the APs distribution by each frequency range and also the overall proportion of the APs over the set of studied proteins. Our analysis showed that the highest concentration of APs occurred in the frequency range of 80-90%, which was confirmed by the binomial test (Figure 3C). These data reinforce our analysis of the hypergeometric (Figure 3A).

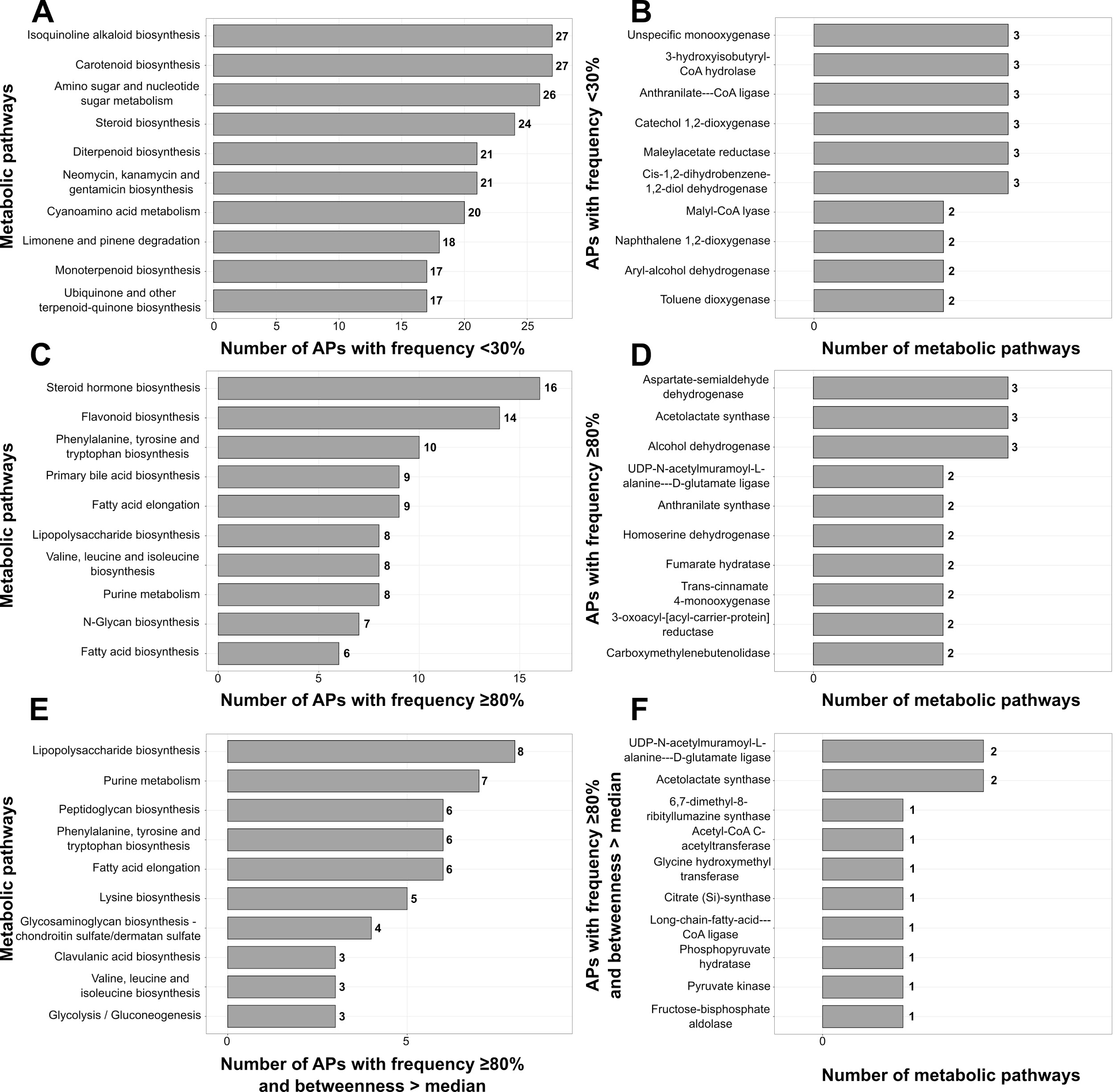
****

**Figure 3: A)** Heatmap showing the concentration of proteins by frequency. The proteins were divided into 2 groups: AP and Non-AP protein. The darker shades of blue indicate a higher concentration of proteins. The lighter shades of blue indicate lower concentrations of proteins. The dashed line represents the cut-off point of the hypergeometric. Proteins that fall below the cut-off point are randomly distributed when compared to proteins above the cut of the hypergeometric. **B)** Betweenness in APs frequencies groups <30% and >=80%. Distribution of betweenness in proteins groups lower than 30% (853) or greater/equal than 80% (206) of frequency. Analyses were performed using 1059 APs. Upper and lower whiskers (dashed line in blue) mark Q1-1.5 x IQR (interquartile range) and Q3+1.5 x IQR points, respectively. Points above and below those marks are considered outliers. The red line represents the median, and the red dot the mean of the entire distribution. The p-value assigned refers to a Wilcoxon-Mann-Whitney test, comparing both distributions. **C)** Ratio of APs by protein frequency ranges. The bars correspond to APs over total proteins found in each frequency range. Each bin represents an interval of 10%. The blue bar corresponds to the interval where the proportion of APs is considered significant (Binomial test, p-value<0.01), which p-value is highlighted. The dashed line represents the expected ratio of APs in the entire protein dataset (20.46%).

## Functional analysis

To suggest the main metabolic pathways in which the APs are involved, we performed an analysis using the KEGG database and the APs studied in our work (Figure 4). We observed that aspartate-semialdehyde dehydrogenase and alcohol dehydrogenase, enzymes that catalyze the electrons transfer from the reductant to the oxidant (electron acceptor), were the APs associated with a greater amount of metabolic pathways in this work (6 metabolic pathways). In fact, 351 types of oxidoreductase can be found in the DEG database (http://www.essentialgene.org/), a database about essential GENES, demonstrating the essentiality of this AP class. In this context, 132 oxidoreductase types are expressed in bacterias, 9 oxidoreductase types are expressed in archaea, and 210 oxidoreductase types are expressed in eukaryotes.

Additionally, we could observe that the steroid biosynthesis is the metabolic pathway with the highest number of articulation points with frequency higher than 80% (16 APs). Moreover, Isoquinoline biosynthesis and carotenoid biosynthesis are the metabolic pathways with the highest number of articulation points with frequency lower than 30% (27 APs). Besides, we evaluated the functions and pathways of the APs associated with the greatest number of pathways. Aminoacid biosynthesis is the main function related to these APs (Table 1). In fact, cysteine and methionine metabolism, as well as Glycine, serine and threonine metabolism are the main pathways related to these APs (Table 1).



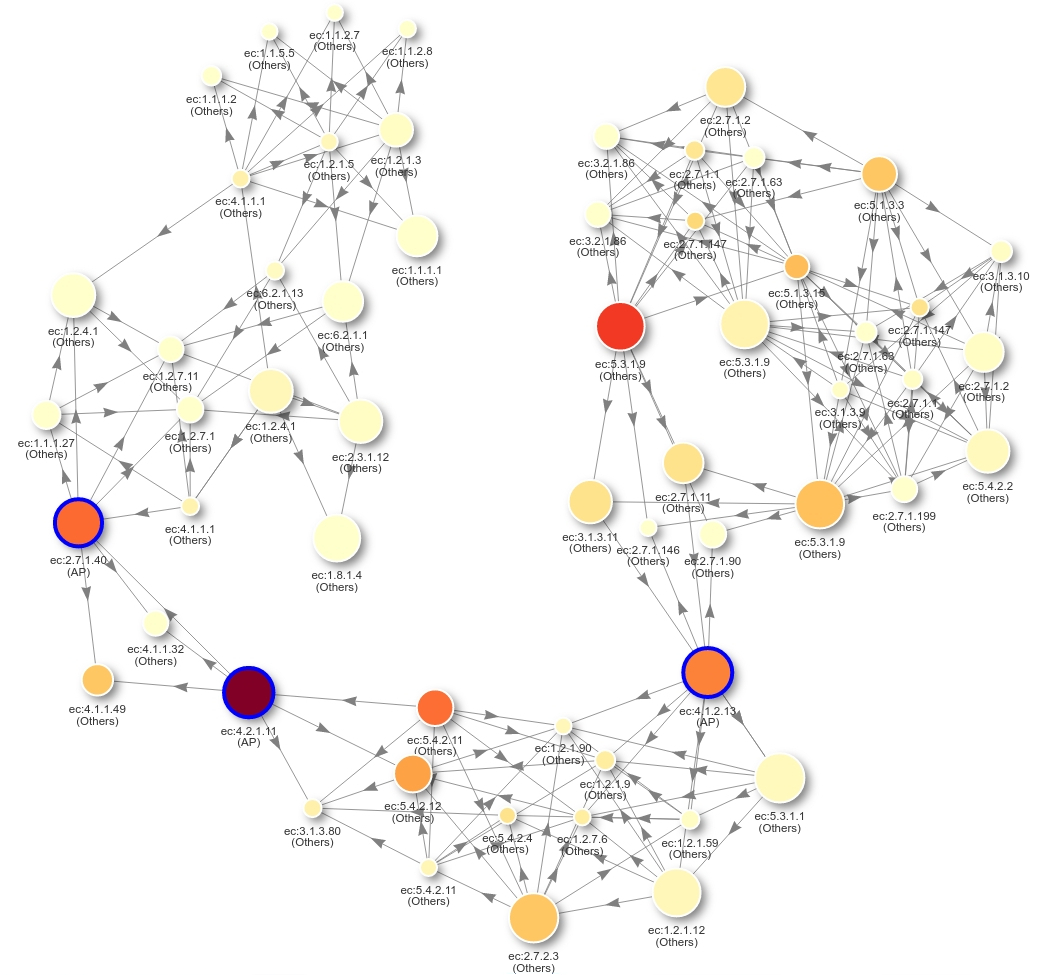
**Figure 4:** A) Top 10 metabolic pathways with the greatest number of APs with a frequency lower than 30% B) Top 10 AP enzymes with a frequency lower than 30% present in the highest number of metabolic pathways. C) Top 10 metabolic pathways with the greatest number of APs with a frequency greater/equal than 80%. D) Top 10 AP enzymes with a frequency greater/equal than 80% present in the highest number of metabolic pathways.

|  |  |  |
| --- | --- | --- |
| **AP, class, and average impact** | **Metabolic Pathways** | **Functions in prokaryotes** |
| * EC 1.2.1.11 - Aspartate-semialdehyde dehydrogenase * Oxidoreductase * 1 | 1. Glycine, serine and threonine metabolism  2. Cysteine and methionine metabolism  3. Lysine biosynthesis | Biosynthesis of amino acids in prokaryotes, fungi, and some higher plants. It forms an early branch point in the metabolic pathway forming lysine, methionine, leucine and isoleucine from aspartate. |
| * EC 2.2.1.6 - Acetolactate synthase * Transferases * 1 | 1. Valine, leucine and isoleucine biosynthesis  2. Butanoate metabolism  3. C5-Branched dibasic acid metabolism | Catalyzes the first step in the synthesis of the branched-chain amino acids in plants, animals, and microorganisms, transfering aldehyde or ketone residues. |
| * EC 1.1.1.1 - Alcohol dehydrogenase * Oxidoreductase * 1 | 1. Chloroalkane and chloroalkane degradation  2. Metabolism of xenobiotics by cytochrome P450  3. Tyrosine metabolism | Facilitate the interconversion between alcohols and aldehydes or ketones with the reduction of nicotinamide adenine dinucleotide (NAD+) to NADH. In yeast, plants, and many bacteria, some alcohol dehydrogenases catalyze the opposite reaction as part of fermentation to ensure a constant supply of NAD+. In humans and animals, they serve to break down alcohols that otherwise are toxic, and they also participate in generation of useful aldehyde, ketone, or alcohol groups during biosynthesis of various metabolites. |
| * EC 6.3.2.9 - D-glutamate ligase * Ligases * 3 | 1. D-Glutamine and D-glutamate metabolism  2. Peptidoglycan biosynthesis | Form carbon-nitrogen bonds as acid-D-amino-acid ligases (peptide synthases). |
| * EC 4.1.3.27 - Anthranilate synthase * Lyases * 4 | 1. Phenylalanine, tyrosine and tryptophan biosynthesis  2. Phenazine biosynthesis | Catalyzes the two-step biosynthesis of anthranilate, an intermediate in the biosynthesis of L-tryptophan. In the first step, the glutamine-binding beta subunit (TrpG) of anthranilate synthase (AS) provides the glutamine amidotransferase activity which generates ammonia as a substrate that, along with chorismate, is used in the second step, catalyzed by the large alpha subunit of AS (TrpE) to produce anthranilate. In the absence of TrpG, TrpE can synthesize anthranilate directly from chorismate and high concentrations of ammonia. |
| * EC 1.1.1.3 - Homoserine dehydrogenase * Oxidoreductase * 2 | 1. Glycine, serine and threonine metabolism  2. Cysteine and methionine metabolism | Catalyses the third step in the aspartate pathway; the NAD(P)-dependent reduction of aspartate beta-semialdehyde into homoserine. Homoserine is an intermediate in the biosynthesis of threonine, isoleucine, and methionine. In photosynthetic organisms, glutamine, glutamate, and aspartate accumulate during the day and are used to synthesize other amino acids. At night, aspartate is converted to asparagine for storage. Mammals lack the enzymes involved in the aspartate metabolic pathway, including homoserine dehydrogenase. As lysine, threonine, methionine, and isoleucine are made in this pathway, they are considered essential amino acids for mammals. |
| * EC 4.2.1.2 - Fumarate hydratase * Lyases * 6 | 1. Citrate cycle (TCA cycle)  2. Pyruvate metabolism | Catalyzes the reversible hydration/dehydration of fumarate to malate. Fumarase comes in two forms: mitochondrial and cytosolic. The mitochondrial isoenzyme is involved in the Krebs Cycle and the cytosolic isoenzyme is involved in the metabolism of amino acids and fumarate. |
| * EC 1.14.14.91 - Cinnamate 4-monooxygenase * Oxidoreductases * 1 | 1. Flavonoid biosynthesis  2. Stilbenoid, diarylheptanoid and gingerol biosynthesis | Acts on paired donors, with O2 as oxidant and incorporation or reduction of oxygen. The oxygen incorporated need not be derived from O2 with NADH or NADPH as one donor, and incorporation of one oxygen atom into the other donor. |
| * EC 1.1.1.100 - Beta-ketoacyl reductase * Oxidoreductases * 2 | 1. Fatty acid biosynthesis  2. Biotin metabolism | Catalyzes the reaction of conversion of 3-hydroxyacyl-[acyl-carrier-protein] to 3-oxoacyl-[acyl-carrier-protein] using NADPH as a reducing agent. Acts on the CH-OH group as hydride donors with NAD+ or NADP+ as hydride acceptors. |
| * 3.1.1.45 - Carboxyme thylenebutenolidase * Hydrolases * 2 | 1. Fluorobenzoate degradation  2. Toluene degradation | Acts on carboxylic ester bonds. |

**Table 1:** Functions and metabolic pathways related to the APs present in a higher number of metabolic pathways. The metabolic pathways were obtained from KEGG, the APs functions were obtained from Uniprot.

## Network visualization

The network elaborated in our study (Figure 5) contemplates protein classifications according to the AP detection algorithm, facilitating the visual identification of the most important proteins of a network (points of articulation). Additionally, our proposal provides dynamic HTML visualizations available at https://igorabrandao.com.br/kegg-pathway-bottleneck, which allow the observation of several metabolic pathways. It is possible to adjust the network size and position, as well as to select proteins according to their classification, making it possible to highlight protein groups of interest. Finally, the AP remotion can disconnect the network and the proteins metrics (impact, degree, and betweenness) can be easily viewed.



**Figure 5:** Visualization of pathway 0010 - glycolysis and gluconeogenesis. Dynamic visualization generated with 58 nodes and 190 edges. Nodes represent enzymes and edges represent reactions. Node size represents protein frequency, colored borders indicate APs, color in the center of node indicates betweenness level. The classification and EC code of the protein are present on the node label. The network was made using the R programming language (visNetwork package).

# 

# DISCUSSION

The metabolic pathway assessment can be approached from the top to the bottom, starting from the network’s topological properties and graphical disposal and moving to the proteins' specific functions, and how they interact within the set as a whole. The present study aimed to investigate the APs for all metabolic pathways available in the KEGG database. To achieve this aim, a pipeline based on R programming language was developed and applied to KEGG XML files for each species (instance) related to each metabolic pathway. We observed that approximately 32% of metabolic pathways are related to less than 16% of KEGG species. This might be due to the existence of specific pathways for certain plants, bacteria, and fungi, which are involved with their own cellular functions in these organisms. An example would be Benzoxazinoid biosynthesis route 00402 (Supporting material S2a) responsible for the defense of some plant species (Supporting material S2b). Also, approximately 31% of metabolic pathways are related to at least 83% of KEGG species, this may be due to these pathways being essential in different types of living beings (Supporting material S3), such as glycolysis and gluconeogenesis route 00010, present in 6084 species (Supporting material S4).

Besides, we studied the profile of each protein within the metabolic pathways. AP represented 20.46% of the proteins studied (Figure 2). This may be due to the topology of metabolic pathways, which are made up of several protein communities that connect through a few APs. The remotion of one of these APs will cause the network rupture since it is a central point that connects various network complexes and/or peripheral regions. Therefore we can suggest that APs might be essential proteins. Other proteins correspond to 79.54% of the proteins because, in terms of network topology, they do not play a crucial role in terms of topology and are generally located within the communities and peripheries of the network.

It is possible to evaluate the metabolic pathway profile through the protein types distribution. It is noteworthy that some metabolic pathways were composed mostly of APs, such as Benzoxazinoid biosynthesis pathway 00402 (Supporting Information S2), which is composed of only eight proteins, of which six are APs. Besides, it is formed by a one-way bridge in terms of mass flow. Other pathways had just peripheral APs, such as caffeine metabolism pathway 00232 (Supporting Information S5). In this case, there are few communities, which are overcrowded and some of them are connected through bridges that represent one-way paths. These bridges are composed of APs, which connect with only one protein. Finally, we highlight the 00061 pathway (Supporting information S5), related to fatty acid synthesis. This pathway is made up of an overcrowded central community, 7 peripheries highly connected to the center and a periphery connected to the center through just one bridge. In addition, this pathway has 3 components disconnected from the main network component, with their own APs. However, although there are several highly connected nodes in the network central area, none of them are an AP due to the connection handles, which guarantee redundancy in this area of the network. In addition to the network topological analysis, it is important to understand the profile of proteins classified as AP. For this purpose, we used hypergeometric and binomial statistical analysis.

The hypergeometric analysis might be used to identify groups of APs that are more relevant. In the hypergeometric analysis, we found a considerable concentration of APs in the protein groups with higher frequencies (Figure 3). We verified that APs tend to be high-frequency proteins, suggesting that these proteins may be more essential to the metabolic pathway when compared with proteins with lower-frequency. However, we found that APs can belong to both groups of high and low frequencies. For this reason, we studied the APs topology using the betweenness metric which identified that the groups of APs with higher frequencies are located in more central regions compared to the groups of APs with lower frequencies. In practice, this implies that APs with low betweenness disconnect peripheral regions from the network, resulting in a disbalance between the components of the remaining network, while the APs with high betweenness are probably located in central regions. Therefore, when these APs are removed from the network, they disconnect it in more homogeneous components.

In the 1047 proteins with higher frequencies (19.04% of total proteins studied), we found a high concentration of APs (236), suggesting that they were strongly present in most species on different metabolic pathways. The other 81% of the studied protein had AP randomly distributed across the frequency groups. The low frequency may be related to the fact that these APs are connectors in peripheral regions (Figure 2C) of the metabolic pathways, being responsible only for connecting a few isolated proteins. Besides, we observed that APs had the highest impact and frequency rates on the networks. It suggests that this set of proteins is essential for its metabolic pathways since they are present in most of the species containing a particular metabolic pathway (Figure 4).

Several studies used hypergeometric as a mathematical model to estimate the data accuracy. For example, a hypergeometric was applied to extract the significant drug-diagnosis associations (based on Bonferroni-adjusted hypergeometric) using a computational method based on Electronic Medical Record (EMR) datasets. This allowed finding differences between pediatric and adult drug use to be compared based on different EMR datasets [23]. Besides, hypergeometric was applied using the physical interactions of transcriptional regulators and protein-protein interaction networks (PPI) to determine the main biomolecules involved in rheumatoid arthritis [24].

Interestingly, a study evaluated the gene co-expression networks for 17 bacterial organisms from the COLOMBOS database performing a weighted gene co-expression network analysis and clustered into modules of genes with similar expression patterns for each species. These networks were analyzed to determine relevant modules through a hypergeometric approach based on a set of transcription factors and enzymes for each genome. The richest modules were characterized using PFAM families and KEGG metabolic maps. Additionally, a GO analysis for enrichment of biological functions was conducted, identifying modules that shared similarity through all the studied organisms by using comparative genomics [25].

Furthermore, Huang et al. intensified maladjusted genes shared between hepatitis B virus and hepatocellular carcinoma by disease-related differentially expressed genes. The PPI network based on dysfunctional genes found dysfunctional modules and significant crosstalk between modules based on the hypergeometric test [26]. Finally, hypergeometric is widely used to test the enrichment of the differentially expressed genes in GO terms or KEGG pathways, especially related to cancer and to verify potential regulation of multi brain factors [27-29].

Biologically, essential genes refer to a group of fundamental genes necessary for a specific organism to survive in a specific environment [30]. The applications of essentiality are varied, ranging from finding the minimal genome required for sustenance to ranking drug targets. Essential genes have been experimentally identified using transposon mutagenesis, antisense RNA, RNA interference, and single-gene deletion. However, experimental determination of essential genes is expensive, time-consuming, and laborious. Therefore, computational predictions of essential genes can give a prioritized shortlist for experimental validation [31].

A study demonstrated that an important protein in a network can be calculated based on the betweenness centrality metric plus its degrees [10]. Other studies suggest performing the identification and removal of APs to provide a new perspective on the organizational principles of complex networks [32]. Finally, to simplify the understanding of the metabolic pathways’ topology, our work proposes the visualization of these pathways as dynamic networks (figure 5). Network visualization is an important aspect of this work since the dynamic visualization provides the possibility to explore network topological features graphically, providing rich details, facilitating the network understanding, and extracting relevant biological information [8].

One study related to inflammatory bowel disease (IBD) used protein-protein interaction network analysis. This study found that there are seven hub-bottleneck proteins in the IBD network responsible to maintain the network integrity. Network evaluation and complex analysis of IBD essential proteins are important to provide a new glance of the disease. Proteins are in a complex interactome organization that any small changes in each individual may lead to dysfunction of the whole system [33].

Studies have shown that network integrity can be disrupted, contributing to serious impairments in disorders such as depression, psychotic disorders, and motor diseases. Investigations into the integration and coordination of large-scale functional networks subserving various conditions must be explored. For example, assessing non-static functional connectivity within whole-brain networks can be important to unravel some mysteries in the central nervous system. In fact, studies assessing network integrity in the central nervous system were developed. However, it is also crucial to identify essential proteins in these networks [34-36].

Given the functional interdependencies between the molecular components in a human cell, a disease is rarely a consequence of an abnormality in a single gene but reflects the perturbations of the complex intracellular and intercellular network that links tissue and organ systems. The emerging tools of network medicine offer a platform to systematically explore not only the molecular complexity of a particular disease, leading to the identification of disease modules and pathways, but also the molecular relationships among apparently distinct (patho)phenotypes. Advances are essential for identifying new disease genes, for uncovering the biological significance of disease-associated mutations identified by genome-wide association studies and full-genome sequencing, and for identifying drug targets and biomarkers for complex diseases [37].

Through several methodologies, such as microarrays, network analysis, differential expression analysis, meta-analysis, functional analysis, database validations, among others, network-based biomarkers have been proposed for various conditions, especially for cancer [38-40]. The network-based biomarker consists of two protein association networks constructed for disease samples and non-disease samples [41]. Network-based biomarkers can be associated with animal behavior models, in which a change in behavior resulting from a manipulation of a network biomarker would constitute a strong validation of drivers of symptoms [42].

To avoid the laborious manual curation for network construction, some methods are developed to automatically reconstruct networks by retrieving interactions or sub-networks from existing maps and models [43, 44]. Combining automatic reconstruction with domain knowledge, a manual search of literature and databases would provide a reasonable strategy to construct detailed and fully annotated large-scale biochemical networks [45]. In this sense, functional analysis can be an interesting tool to assess the functions most associated with specific protein classes, for example, APs. However, establishing these correspondences between different databases (e.g., KEGG, BioCyc, and Gene Ontology) is a non-trivial problem because of the non-standard terminology used in the scientific literature (e.g., large numbers of synonyms are used for a given compound) [46]. For this, DAVID and KEGG mapper software can be used to obtain the protein identity to perform functional analysis.

We performed a functional analysis to determine the main metabolic pathways regarding the APs using KEGG ECs. A study also used KEGG to perform functional analyses and studies used Gene Ontology to determine the biological process and molecular functions of proteins related to cancer [47, 48]. However, it is important to highlight that no other study evaluated the main metabolic pathways associated with all APs. Most APs with the greatest number of associated pathways are oxidoreductases (AP classes) and the synthesis of steroids was the metabolic pathway with the highest number of APs with a frequency higher than 80%. In fact, on 00140 pathway (Steroid hormone biosynthesis) out of 16 total APs, 14 are oxidoreductases. Oxidoreductases participate in the metabolism of some ethers. A 5α-reductase (5α-R) and 3α- or 3β-hydroxysteroid oxidoreductase (HSOR), is important to the generation of reduced metabolites of progesterone and testosterone, acting on several steroid and neurotransmitter receptors and regulating the neural function and behavior.

Progesterone is metabolized into dihydro-progesterone by the action of 5α-R and subsequently into 3α, 5α-tetrahydro-progesterone, or into 3β, 5α-tetrahydro-progesterone, by the action of 3α- or 3β-HSOR, respectively. Besides, testosterone is converted into dihydrotestosterone and subsequently into 5α-androstane-3α,17β-diol or 5α-androstane-3β,17β-diol. The enzymatic reaction performed by 5α-R is the rate-limiting step during the 3α,5αreduced synthesis. The Oxidoreductase can also convert corticosterone and deoxycorticosterone into dihydro-corticosteroids and dihydro-deoxycorticosterone, respectively, and this latter metabolite can be converted into 3α,5α-tetrahydro-desoxycorticosterone [49].

In summary, we suggested that bottlenecks are articulation points with highest frequencies and located in the center of the network. New studies will be required to establish more relationships between these key proteins (APs) and topological and biological attributes of the metabolic pathways. Additional studies with metabolomics, which is a powerful technology that allows for the assessment of global metabolic profiles, can be used to distinguish between diseased and non-diseased status information [50]. Another future possibility is the usage of machine learning (ML) algorithms to create a predictive model to identify potential essential proteins in various metabolic pathways. Finally, knockout studies can be applied to experimentally evaluate whether specific APs are essential for a given metabolic pathway. For example, recently a receptor involved in memory (NOP receptor) has been experimentally associated with metabolic pathways involved in depression; it remains to assess this receptor essentiality on the path through bioinformatics tools.

# 

# CONCLUSION

We found that 20.46% of the proteins are articulation points (1125). The most important articulation points are high-frequency proteins which are located in central regions of the network. Moreover, the highest concentration of articulation points occurred in the frequency range between 80% and 90% (p < 0.01). A pattern of non-randomness of articulation points was identified in the protein groups that have a frequency of at least 74.5% (p < 0.01). Most APs with frequency higher/equal than 80% are related to steroid biosynthesis metabolism. Besides, within this group, the oxidoreductase are the APs with the highest number of related metabolic pathways. It remains to perform a deep analysis on the articulation points biological roles. This work contributes to the study of metabolic pathways using computational approaches since nowadays few works are exploring massive data related to curated databases of metabolic pathways and generating analysis to help to understand the big picture of metabolic pathways.

# REFERENCES

1. Jeong H, Mason SP, Barabási AL, Oltvai ZN. Lethality and centrality in protein networks. Nature 2001; 411(6833): 41.

2. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nature reviews genetics 2004; 5(2): 101.

3. Larsen TS. The Scientist's Guide to Cardiac Metabolism: Historical Perspectives (cap.15). 1st ed. Germany: Academic Press; 2016: 207-217.

4. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. Hierarchical organization of modularity in metabolic networks. Science 2002; 297(5586): 1551-1555.

5. Ma HW, Zeng AP. The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics 2003; 19(11): 1423-1430.

6. Maniadi EM, Tollis IG. Analysis and visualization of metabolic pathways and networks: A hypergraph approach. 1st ed. Spain: Valencia; 2014: 109-112.

7. Mahadevan R, Palsson BO. Properties of metabolic networks: structure versus function. Biophysical journal 2005; 88(1): 7-9.

8. Beguerisse-Díaz M, Bosque G, Oyarzún D, Picó J, Barahona M. Flux-dependent graphs for metabolic networks. NPJ systems biology and applications 2018; 4(1): 32.

9. Ferreira RM, Rybarczyk-Filho JL, Dalmolin RJ, Castro MA, Moreira JC, Brunnet LG, de Almeida RM. Preferential duplication of intermodular hub genes: an evolutionary signature in eukaryotes genome networks. PloS one 2013; 8(2).

10. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS computational biology 2007; 3(4): e59.

11. Behzad M, Chartrand G. Introduction to the Theory of Graphs. 6st ed. Allyn and Bacon. EUA: Boston; 1972.

12. Harary F. Graph theory. 6st ed. Addison-Wesley. England: London; 1996.

13. Ausiello G, Firmani D, Laura AL. Real‐time monitoring of undirected networks: Articulation points, bridges, and connected and biconnected components. Networks 2012; 59(3): 275-288.

14. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic acids research 2016; 45(1): 353-361.

15. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. Nature 2000; 407(6804): 651.

16. Wagner A, Fell DA. The small world inside large metabolic networks. Proceedings of the Royal Society of London 2001; 268(1478): 1803-1810.

17. Ma HW, Zeng AP. The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics 2003; 19(11): 1423-1430.

18. Farina G. A linear time algorithm to compute the impact of all the articulation points 2015; 1: 1-4.

19. Tenenbaum D. KEGGREST: Client-side REST access to KEGG. R package version 2016; 1(1).

20. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal, Complex Systems 2006; 1695(5): 1-9.

21. Hopcroft J, Tarjan R. Algorithm 447: efficient algorithms for graph manipulation. Communications of the ACM 1973; 16(6): 372-378.

22. Almende BV, Thieurmel B Robert T. visNetwork: Network Visualization using ‘vis. js’ Library, R package version 2016; 1(1).

23. Yu G, Zeng X, Ni S, Jia Z, Chen W, Lu X et al. A computational method to quantitatively measure pediatric drug safety using electronic medical records. BMC Medical Research Methodology 2020; 20(1): 1-11.

24. Comertpay B, Gov E. Identification of key biomolecules in rheumatoid arthritis through the reconstruction of comprehensive disease-specific biological networks. Autoimmunity 2020: 1-11.

25. Perez-Rueda E, Galan-Vasquez E. Identification of modules with similar gene regulation and metabolic functions based on co-expression data. Frontiers in Molecular Biosciences 2019; 6: 139.

26. Huang XB, He YG, Zheng L, Feng H, Li YM, Li HY et al. Identification of hepatitis B virus and liver cancer bridge molecules based on functional module network. World journal of gastroenterology 2019; 25(33): 4921.

27. Zhang PB, Huang ZL, Xu YH, Huang J, Huang XY, Huang XY. Systematic analysis of gene expression profiles reveals prognostic stratification and underlying mechanisms for muscle-invasive bladder cancer. Cancer Cell International 2019; 19(1): 1-13.

28. Varallo GR, Jardim-Perassi BV, Alexandre PA, Fukumasu H, Zuccari DAPC. Global gene expression profile in canine mammary carcinomas. The Veterinary Journal 2019; 254: 105393.

29. Wang Z, Du X, Yang Y, Zhang G. Study on miR-384-5p activates TGF-β signaling pathway to promote neuronal damage in abutment nucleus of rats based on deep learning. Artificial Intelligence in Medicine 2019; 101: 101740.

30. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G et al. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. Cell 2015; 163(6): 1515-1526.

31. Azhagesan K, Ravindran B, Raman K. Network-based features enable prediction of essential genes across diverse organisms. PloS one 2018; 13(12).

32. Tian L, Bashan A, Shi D-N, Liu Y-Y. Articulation points in complex networks. Nat Commun 2017; 8(1): 14223.

33. Asadzadeh-Aghdaee H, Shahrokh S, Norouzinia M, Hosseini M, Keramatinia A, Naghibzadeh B et al. Introduction of inflammatory bowel disease biomarkers panel using protein-protein interaction (PPI) network analysis 2016; 9(1): 8-13.

34. Ge R, Downar J, Blumberger DM, Daskalakis ZJ, Lam RW, Vila-Rodriguez F. Structural network integrity of the central executive network is associated with the therapeutic effect of rTMS in treatment resistant depression. Progress in Neuro-Psychopharmacology and Biological Psychiatry 2019; 92: 217-225.

35. Larivière S, Ward NS, Boudrias MH. Disrupted functional network integrity and flexibility after stroke: Relation to motor impairments. NeuroImage: Clinical 2018; 19: 883-891.

36. Sheffield JM, Kandala S, Tamminga CA, Pearlson GD, Keshavan MS, Sweeney JA et al. Transdiagnostic associations between functional brain network integrity and cognition. JAMA psychiatry 2017; 74(6): 605-613.

37. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature reviews genetics 2011; 12(1): 56-68.

38. Zhao H, Li H. Network-based meta-analysis in the identification of biomarkers for papillary thyroid cancer. Gene 2018; 661: 160-168.

39. Wang LX, Li Y, Chen GZ. Network-based co-expression analysis for exploring the potential diagnostic biomarkers of metastatic melanoma. PloS one 2018; 13(1).

40. Zhang Y, Zhu L, Wang X. A Network-Based Approach for Identification of Subtype-Specific Master Regulators in Pancreatic Ductal Adenocarcinoma. Genes 2020; 11(2): 155.

41. Wang X, Gulbahce N, Yu H. Network-based methods for human disease gene prediction. Briefings in Functional Genomics 2011; 10(5): 280–93.

42. Neylan TC, Schadt EE, Yehuda R. Biomarkers for combat-related PTSD: focus on molecular networks from high-dimensional data. European Journal of Psychotraumatology 2014; 5(1): 23938.

43. Ritz A, Poirel CL, Tegge AN, Sharp N, Simmons K, Powell AM et al. Pathways on demand: automated reconstruction of human signaling networks. NPJ systems biology and applications 2016; 2: 16002.

44. Supper J, Spangenberg L, Planatscher H, Dräger A, Schröder A, Zell A. BowTieBuilder: modeling signal transduction pathways BMC Syst Biol. 2009; 3(1): 67.

45. Khan FM, Gupta SK, Wolkenhauer O. Integrative workflows for network analysis. Essays in Biochemistry 2018; 62(4): 549–61.

46. Altman T, Travers M, Kothari A, Caspi R, Karp PD. A systematic comparison of the MetaCyc and KEGG pathway databases. BMC Bioinformatics 2013; 14(1): 112.

47. Wang LX, Li Y, Chen GZ. Network-based co-expression analysis for exploring the potential diagnostic biomarkers of metastatic melanoma. PloS one 2018; 13(1).

48. Perez-Rueda E, Galan-Vasquez E. Identification of modules with similar gene regulation and metabolic functions based on co-expression data. Frontiers in Molecular Biosciences 2019; 6: 139.

49. Giatti S. Diviccaro S, Falvo E, Garcia-Segura LM, Melcangi RC. Physiopathological role of the enzymatic complex 5α-reductase and 3α/β-hydroxysteroid oxidoreductase in the generation of progesterone and testosterone neuroactive metabolites. Frontiers in Neuroendocrinology 2020: 100836.

50. Wang X, Zhang A, Han Y, Wang P, Sun H, Song G et al. Urine Metabolomics Analysis for Biomarker Discovery and Detection of Jaundice Syndrome in Patients With Liver Disease. Mol Cell Proteomics 2012; 11(8): 370–80.

# SUPPORTING INFORMATION

## Supporting information S1: A) APs with high betweenness and frequency of pathway 00010 - glycolysis and gluconeogenesis. I) AP ec:4.1.2.13 with betweenness of 0.0255 and frequency of 96.98%. II) AP ec:4.2.1.11 with betweenness of 0.0490 and frequency of 98.26%. III) AP ec:2.7.1.40 with betweenness of 0.0282 and frequency of 91.73%. B) APs with low betweenness and frequency of pathway 00640 - propanoate metabolism. IV) AP ec:1.1.1.80 with betweenness of 0.0023 and frequency of 2.23%. V) AP ec:4.1.1.4 with betweenness of 0.0034 and frequency of 7.69%. VI) AP ec:2.8.3.8 with betweenness of 0.0042 and frequency of 9.79%. VII) AP ec:1.1.1.77 with betweenness of 0.0017 and frequency of 6.89%.

## 

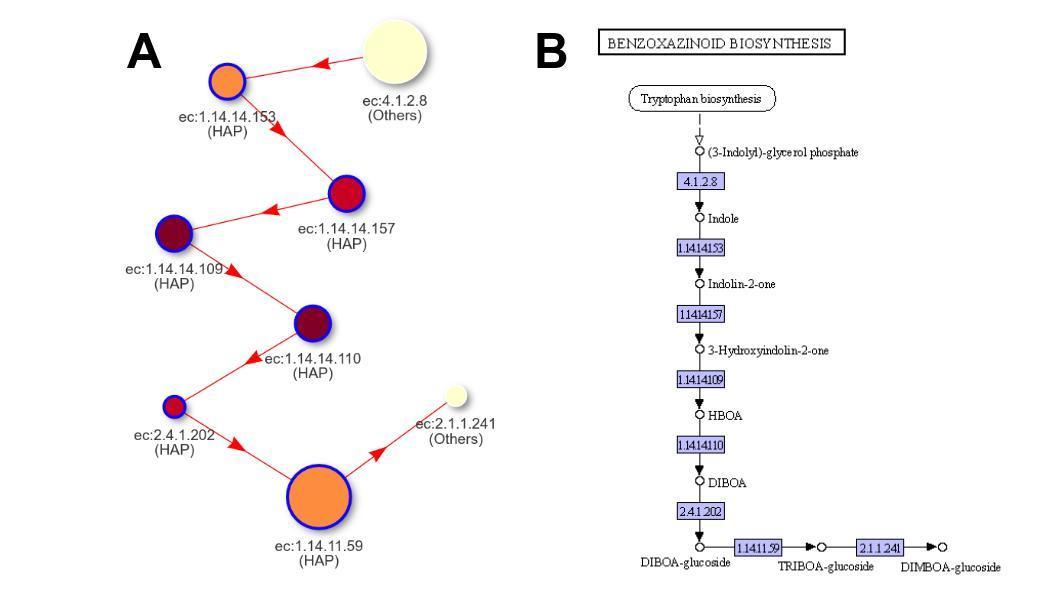
## 

## 

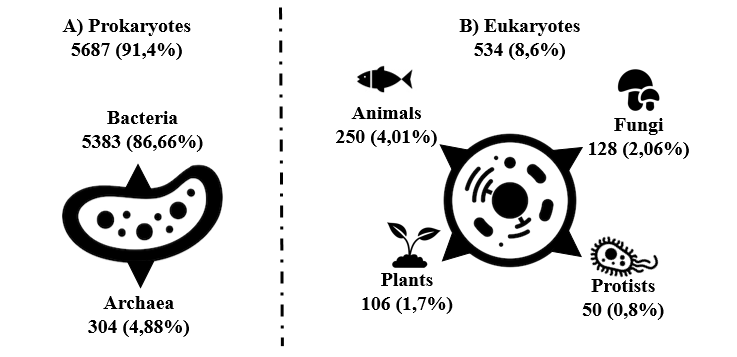
## 

## 

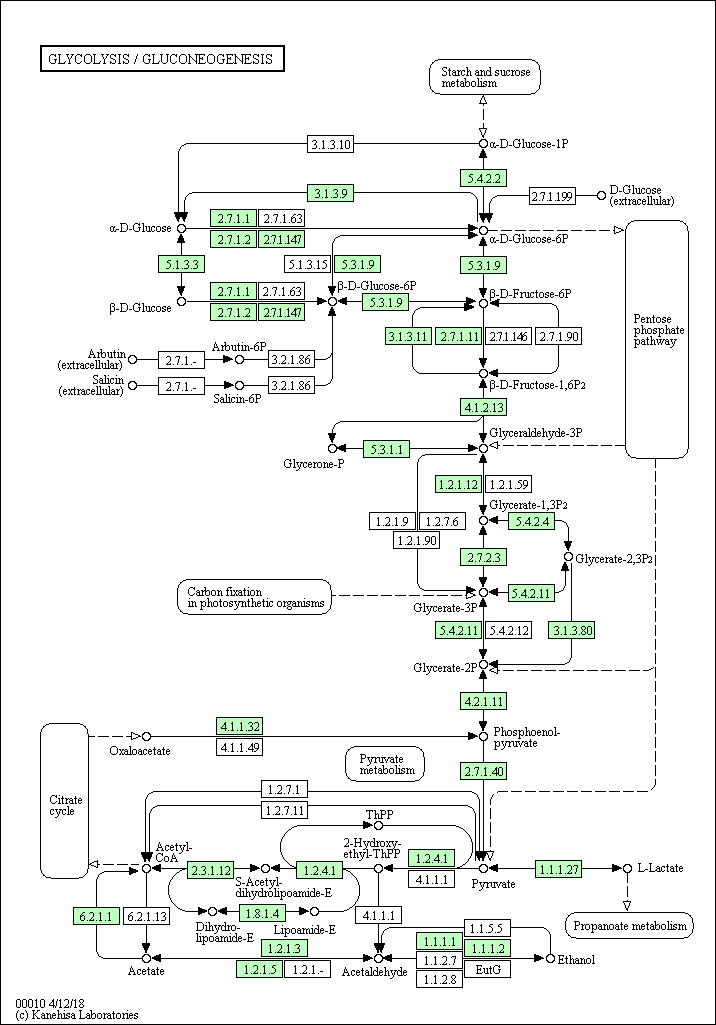
## Supporting information S2: Pathway 00402 - Benzoxazinoid biosynthesis. A) Dynamic visualization generated with 8 nodes and 7 edges. B) KEGG metabolic pathway diagram.



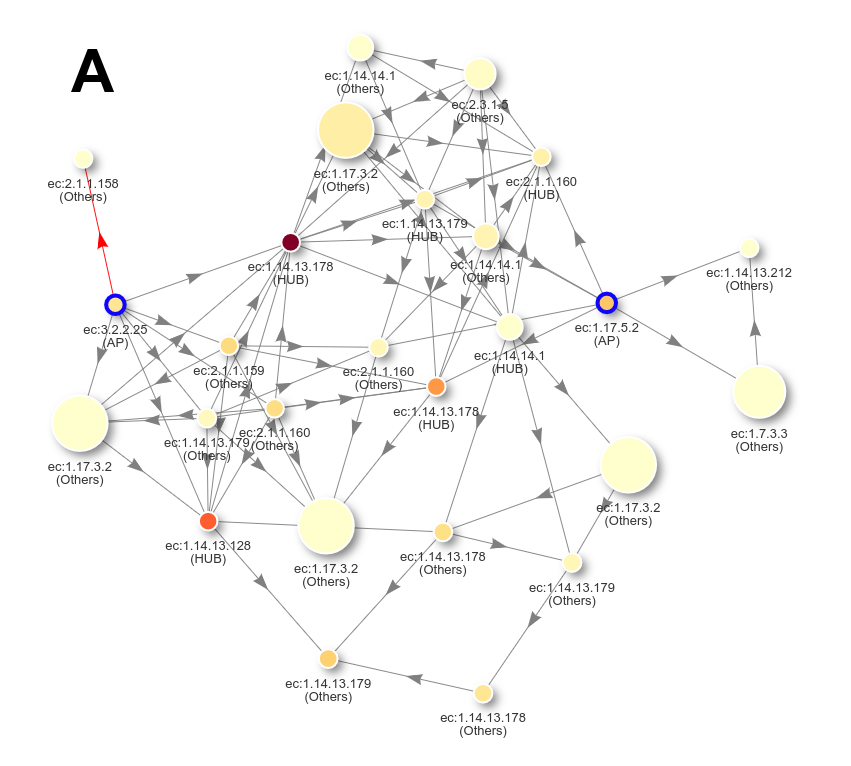
## Supporting information S3: Classification and quantity of organisms used in the study. (A) Prokaryotes; (B) Eukaryotes.

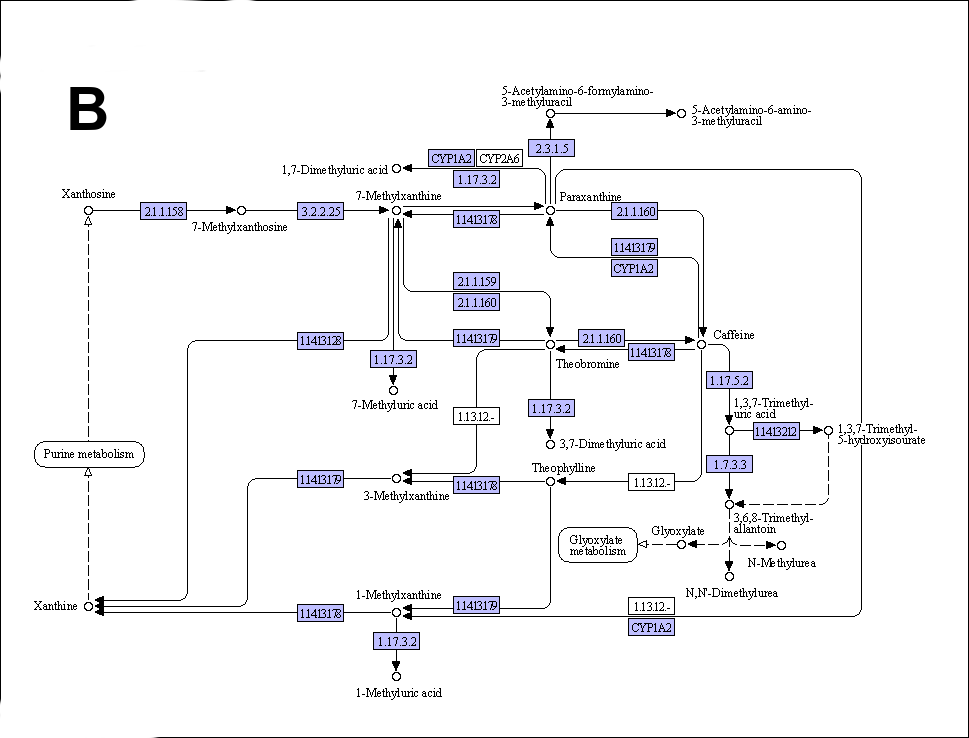


## Supporting information S4: Visualization of pathway 0010 (glycolysis and gluconeogenesis) in KEGG database. KEGG presents information about the metabolic pathway as a sequential block diagram. In which colored or white boxes represent proteins. Colored boxes are proteins present in the metabolic pathway, while whites are absent proteins in these pathways. Source: https://www.kegg.jp/kegg-bin/show\_pathway?org\_name=hsa& mapno=00010.



## Supporting information S5: Pathway 00232 - Caffeine metabolism. A) Dynamic visualization with 26 nodes and 76 edges. B) KEGG metabolic pathway diagram.





## Supporting information S6: Pathway 00061 - Fatty acid biosynthesis. A) Dynamic visualization with 170 nodes and 1063 edges. B) KEGG metabolic pathway diagram.

