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

Running title: INVESTIGATION OF ARTICULATION POINTS

Igor Brandão[1], Diego Arthur[1], Alice Câmara[2], Leonardo Campos[1], Clovis Reis[1], Rodrigo JS Dalmolin[1*]

[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 determines that an essential node in a network is characterized by the articulation point. It remains to be determined whether these articulation points are essential in metabolic pathways and their

topological impact on the network. Using network metrics and articulation points, we look for a reliable way to verify the essentiality of proteins by assessing systematically several metabolic pathways. For this purpose, we determine 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. Most of the articulation points were located in the protein groups with the highest occurrences. Approximately 32% of metabolic pathways are related to 17% of all studied species. Also, we observed that the length of the metabolic networks varies between 2-170 proteins. Proteins classified as Hub articulation points and articulation points represented 20% of the proteins. Most APs are related to amino acid and sugar metabolism and the monooxygenases are the APs with the highest number of related metabolic pathways. The findings suggest that articulation points are proteins with the highest frequencies. This work contributes to the systematic study of metabolic pathways using computational approaches.

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's 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's 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. Since it

represents mass flow, the pathway can be represented by directed graphs. Nodes represent metabolites and edges represent reactions. A directed edge from a compound to a reaction node denotes a reactant while an edge from a reaction to a compound node denotes a product of the reaction [6].

To date, no research systematically evaluated the topology of all metabolic maps as a whole to study the points of articulation. 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, APs can impact the metabolic pathway differently. The impact is defined as the number of vertices that get disconnected from the main (largest) surviving connected component after the removal of the AP [18].

This work aims to identify the APs in KEGG metabolic pathways and calculate the frequency for each protein in each KEGG metabolic pathways, and to evaluate the impact of each AP on the studied metabolic pathways.

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. These lists were the reference for the selection of metabolic pathways used in the study. KGML (KEGG XML) file is processed and transformed into a graph object whose nodes represent the combination of EC numbers (enzyme commission number) with the related chemical reaction code. 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 four groups according to their profile in a given metabolic pathway: HAP (hub articulation point), HP (hub point), AP, and "OTHER" (non-hub non-articulation point). The protein is considered a hub if its degree is within the top 20% of the metabolic pathway [10].

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 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 nodes disconnected represents a higher impact coefficient for an AP and vice-versa. However, a

network 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 [0, 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.

Hypergeometric analyses

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 X times, with X 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 [0, 1]. P-value < 0.05 was considered significant, showing that the AP proteins were not distributed randomly.

Metabolic pathways visualization

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), and protein classification (label) using visNetwork R package [22].

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.

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 all 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 5,000 species.

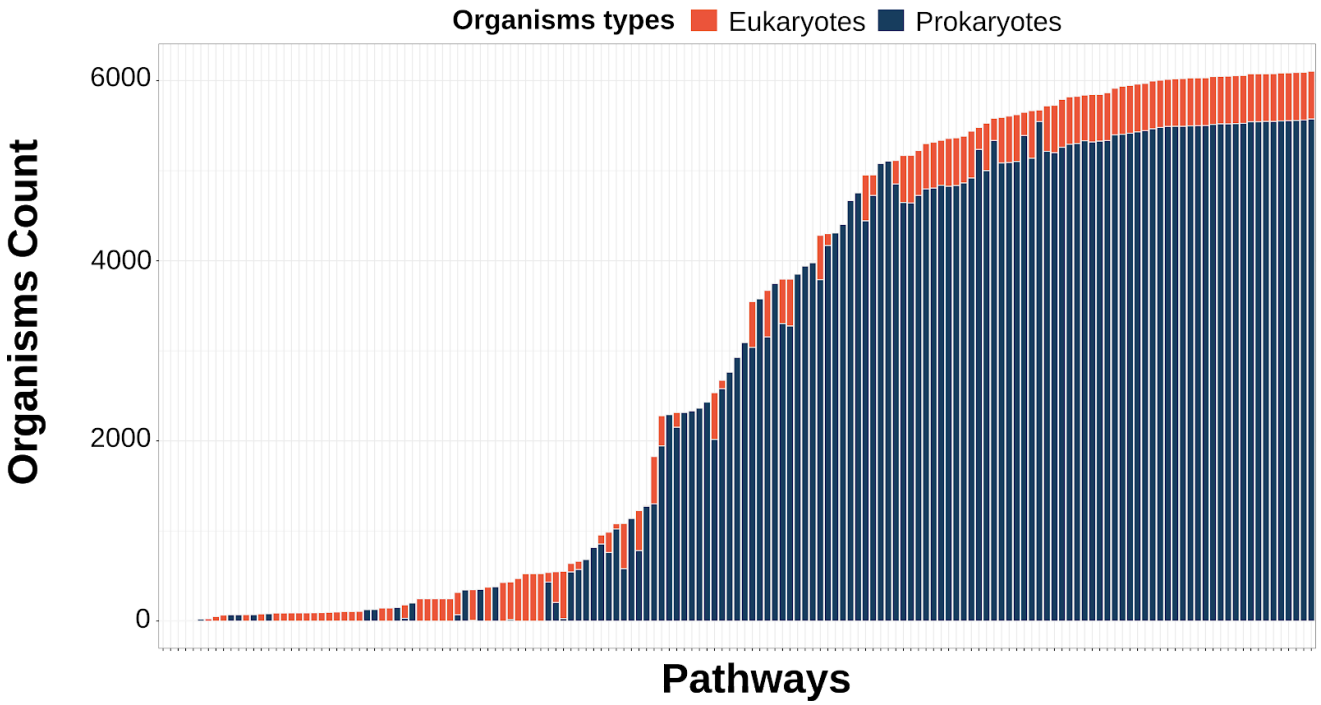


Figure 1: Distribution of species quantity by metabolic pathways available in KEGG. X-axis: metabolic pathway. Y-axis: organisms count. The orange bars represent pathways related to eukaryotes. Eukaryotes represent 8.6% (534) of total species (6221). Blue bars represent pathways related to prokaryotes.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

Protein classification

In the context of the studied networks, each protein is considered a node that is represented by the following entities: enzyme commission number (EC), reaction code,

KEGG (x, y) coordinates. The nodes studied (n=5497) were divided into 4 groups (Figure 2) according to their degree and whether or not it is an AP: HAP (hub articulation point), HP (hub point), and OTHER (non-hub non-articulation point). AP (n=722) represented 13.13% of the nodes studied. Regarding HAP (n=403), a frequency of 7.33% of the analyzed nodes was observed. HUB (n=966) corresponded to approximately 17.57% of total nodes since a few connectors are usually present in a metabolic network. Other nodes (n=3406) corresponded to 61.96% of the total of studied nodes.

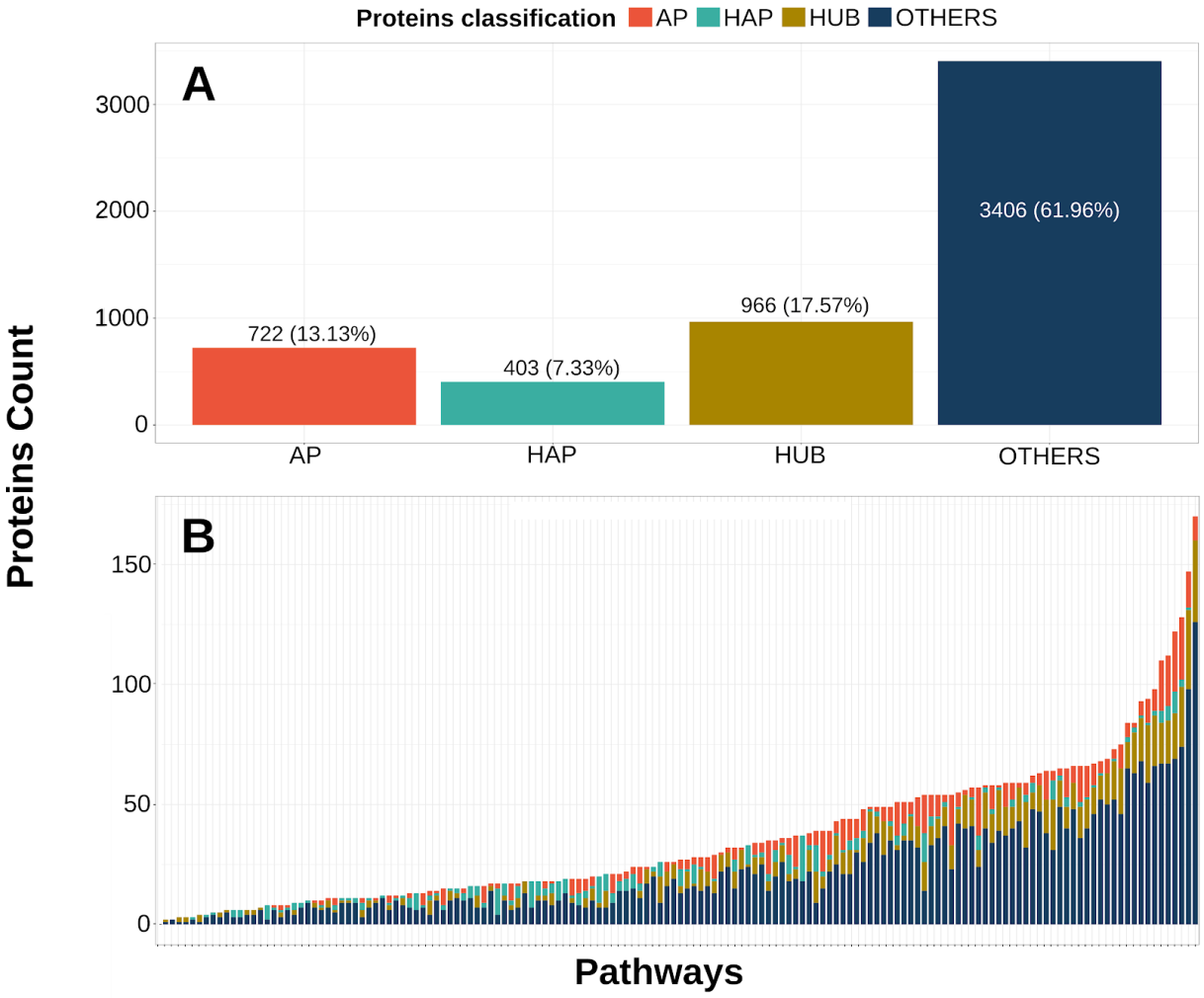


Figure 2: (A) Proteins classifications. AP: articulation point (13.13%); HAP: hub articulation point (7.33%); HP: hub point (17.57%); Others: non-hub non-articulation point (61.96%). Total number of studied nodes: 5497. **(B)** The x-axis represents each pathway and the y-axis

depicts the number of nodes per pathway. Protein count per pathway varies from 2 to 170 nodes.

Hypergeometric analyses

Through the hypergeometric analysis, a pattern of non-randomness was identified in the proteins that have a frequency of at least 74.14%, 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, 1148 more frequent proteins were selected, which presented a higher frequency than the cut-off point, of which 263 are APs (22.9%). About the total of studied proteins, 1125 APs that represent 20.46% were evaluated. Higher AP density was observed above the cut-off point when compared to non-AP proteins (Figure 3a, 3b). Besides, at frequencies between 90 and 93%, there was the highest concentration of AP. On the other hand, there was a great variation below the cut-off point, with frequencies between 40 and 55% standing out, which suggests the randomness of APs in this region (Figure 3c).

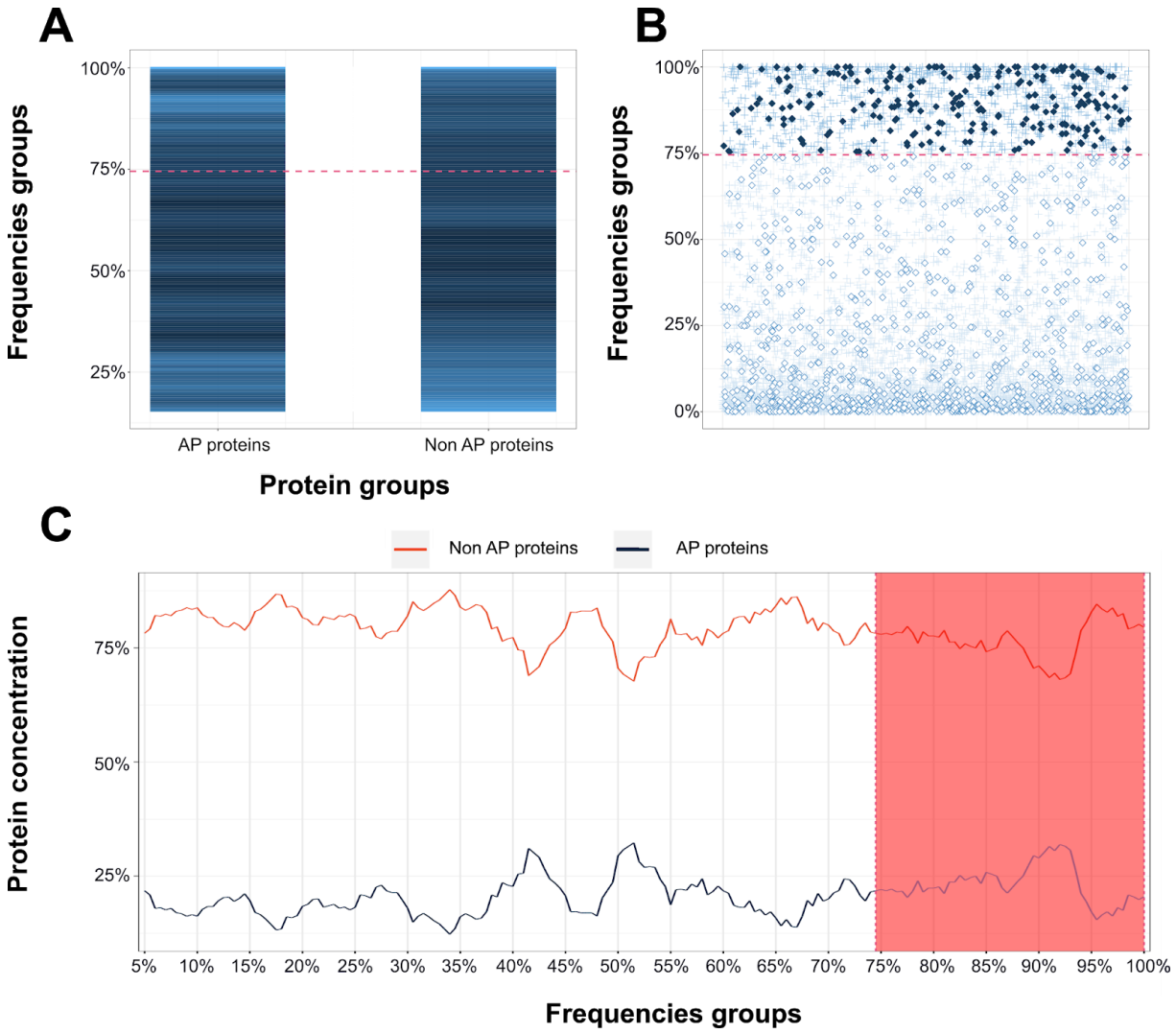


Figure 3: A) Heatmap showing the concentration of proteins by frequency. The proteins were divided into 2 groups: AP and Non-AP protein. The lighter shades of blue indicate a higher concentration of APS. The darker shades of blue indicate lower concentrations of AP. **B)** Protein distribution according to 4 groups: filled diamonds means APs with higher frequencies ($p < 0.05$), empty diamonds mean AP with lower frequencies ($p < 0.05$). Crosses mean Non-APs proteins. **C)** Proteins concentration by frequency. The blue line corresponds to APs and the red line corresponds to Non-APs proteins. The pink area indicates significant proteins (most frequently, $p < 0.05$). The dashed line represents the cut point of the hypergeometric. Proteins that fall below the cut-off point are randomly distributed compared to proteins above the cut of the hypergeometric.

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 monooxygenases, enzymes that incorporate one hydroxyl group into substrates in many metabolic pathways, were the AP type associated with a greater amount of metabolic pathways in this work (10 metabolic pathways). In fact, 41 types of monooxygenases can be found in the DEG database (http://www.essentialgene.org/), a database about essential GENES, demonstrating the essentiality of this AP, mainly in Homo Sapiens, in which 25 types of monooxygenases are expressed.

Additionally, we could observe that the "amino sugar and nucleotide sugar metabolism" pathway was associated with a greater number of APs (35 APs). Besides, we evaluated the functions and the main expression areas of the APs with the greatest impact (impacts between 13 and 38). We observed that the liver and gallbladder were the main human tissues in which most high-impact APs are expressed. Finally, we found that lyases and isomerases were the main classes of the high-impact APs (Table 1).

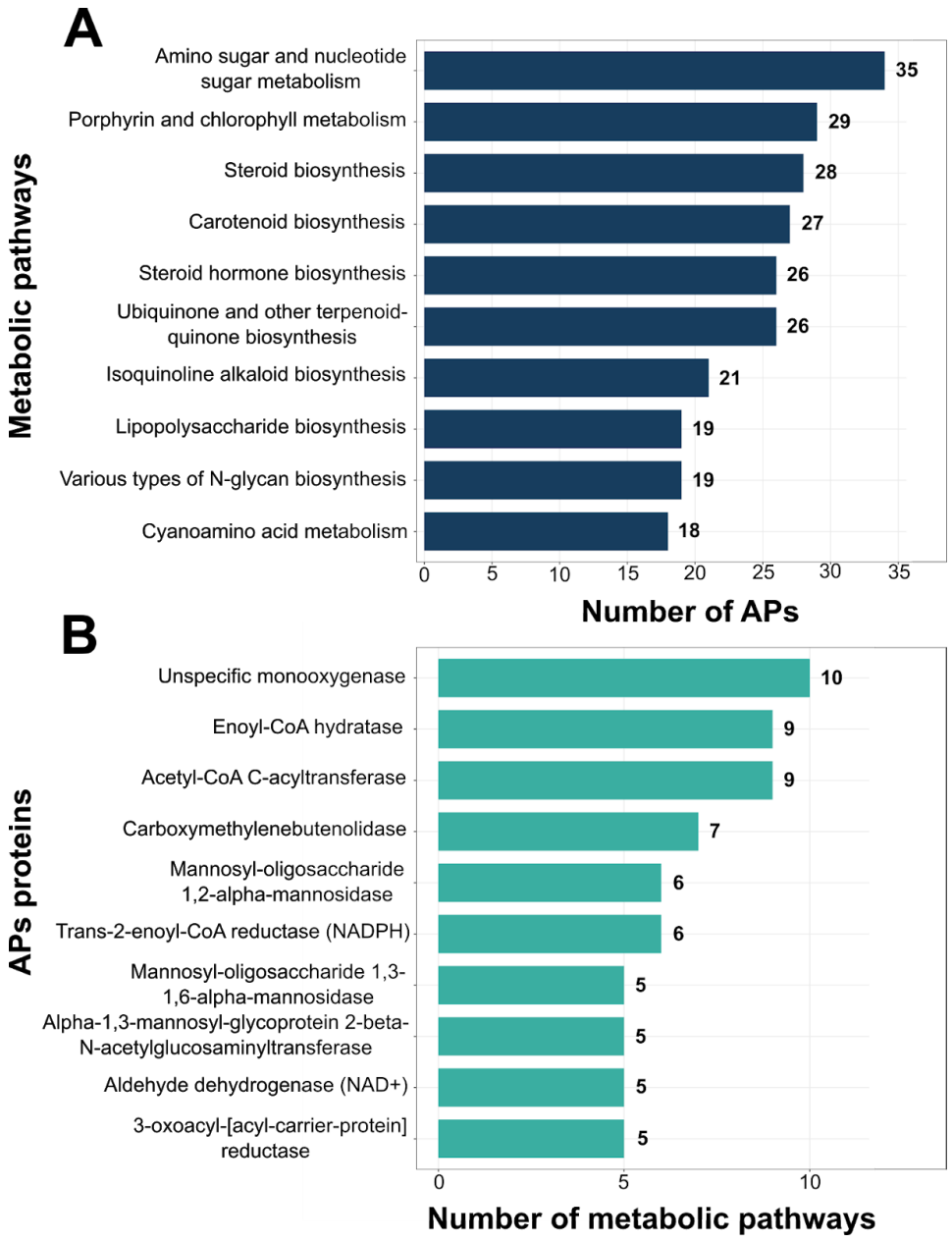


Figure 4: A) Top 10 metabolic pathways with the greatest number of APs. B) Top 10 AP proteins present in a higher number of metabolic pathways.

AP, class, and impact factor	Metabolic Pathways	Functions	Main expression areas
EC 4.4.1.1 Cystathionine gamma-lyase Lyases 38	1. Glycine, serine and threonine metabolism 2. Cysteine and methionine metabolism 3. Selenocompound metabolism 4. Biosynthesis of secondary metabolites	A multifunctional pyridoxal-phosphate protein that converts cystathione derived from methionine into cysteine. The enzyme cleaves a carbon-sulfur bond, releasing L-cysteine and an unstable enamine product that tautomerizes to an imine form, which undergoes hydrolytic deamination to form 2-oxobutanoate and ammonia. The latter reaction can be catalyzed by	Liver & gallbladder

	5. Biosynthesis of antibiotics	2-iminobutanoate/2-iminopropanoate deaminase. Also catalyzes the conversion of L-homoserine to 2-oxobutanoate and ammonia of L-cystine to thiocysteine, pyruvate and ammonia, and of L-cysteine to pyruvate, hydrogen sulfide and ammonia. Glutathione synthesis in the liver is dependent upon the cysteine availability.	
EC 5.4.99.5 Chorismate mutase Isomerases 25	1. Phenylalanine, tyrosine and tryptophan biosynthesis 2. Biosynthesis of secondary metabolites 3. Biosynthesis of antibiotics	Plays an important role in the biosynthesis of aromatic amino acids, catalyzes the first step of the shikimate branch pathway specific to phenylalanine and tyrosine biosynthesis, and catalyzes the Claisen rearrangement of chorismate to prephenate and the decarboxylation/dehydration of prephenate to phenylpyruvate.	Liver & gallbladder
EC 4.1.2.13 Fructose-bisphosphate aldolase Lyases 23	1. Glycolysis / Gluconeogenesis 2. Pentose phosphate pathway 3. Fructose and mannose metabolism 4. Methane metabolism 5. Carbon fixation in photosynthetic organisms 6. Biosynthesis of secondary metabolites 7. Microbial metabolism in diverse environments 8. Biosynthesis of antibiotics	The glycolytic enzyme catalyzes the reversible conversion of fructose-1,6-bisphosphate to glyceraldehyde 3-phosphate and dihydroxyacetone phosphate. Three aldolase isozymes are differentially expressed during development. The developing embryo produces aldolase A. In adult liver, kidney and intestine, aldolase A expression is repressed and aldolase B is produced. In the brain and other nervous tissue, aldolase A and C are expressed equally. There is a high degree of homology between aldolase A and C. Defects in ALDOB cause hereditary fructose intolerance.	Liver & gallbladder
EC 4.2.1.17 Enoyl-CoA hydratase Lyases 16	1. Fatty acid elongation 2. Fatty acid degradation 3. Valine, leucine and isoleucine degradation 4. Geraniol degradation 5. Lysine degradation 6. Phenylalanine metabolism 7. Benzoate degradation 8. Tryptophan metabolism 9. beta-Alanine metabolism 10. alpha-Linolenic acid metabolism 11. Aminobenzoate degradation 12. Propanoate metabolism 13. Butanoate metabolism 14. Carbon fixation pathways 15. Limonene and pinene degradation 16. Caprolactam degradation 17. Biosynthesis of secondary metabolites 18. Microbial metabolism 19. Biosynthesis of antibiotics	The protein is a bifunctional enzyme and is one of the four enzymes of the peroxisomal beta-oxidation pathway. The N-terminal region of the encoded protein contains enoyl-CoA hydratase activity while the C-terminal region contains 3-hydroxyacyl-CoA dehydrogenase activity.	Muscle
EC 4.2.1.2 Fumarate hydratase Lyases 15	1. Citrate cycle (TCA cycle) 2. Pyruvate metabolism 3. Carbon fixation pathways 4. Biosynthesis of secondary metabolites 5. Microbial metabolism 6. Biosynthesis of antibiotics	Catalyzes the reversible stereospecific interconversion of fumarate to L-malate.	Kidney & urinary bladder
EC 5.3.1.9	1. Glycolysis / Gluconeogenesis 2. Pentose phosphate pathway 3. Starch and sucrose metabolism	In the cytoplasm, catalyzes the conversion of glucose-6-phosphate to fructose-6-phosphate, the second step in glycolysis, and the reverse reaction during	Muscle

Glucose-6-phosphate isomerase Isomerases 14	4. Amino sugar and nucleotide sugar metabolism 5. Biosynthesis of secondary metabolites 6. Microbial metabolism 7. Biosynthesis of antibiotics	gluconeogenesis. Besides its role as a glycolytic enzyme, it also acts as a secreted cytokine: acts as an angiogenic factor that stimulates endothelial cell motility. Acts as a neurotrophic factor, neuroleukin, for spinal and sensory neurons. It is secreted by lectin-stimulated T-cells and induces immunoglobulin secretion.	
EC 5.4.2.8 Phosphomannomutase Isomerases 14	1. Fructose and mannose metabolism 2. Amino sugar and nucleotide sugar metabolism 3. Biosynthesis of secondary metabolites 4. Biosynthesis of antibiotics	Phosphomannomutase catalyzes the conversion between D-mannose 6-phosphate and D-mannose 1-phosphate which is a substrate for GDP-mannose synthesis. GDP-mannose is used for the synthesis of dolichol-phosphate-mannose, which is essential for N-linked glycosylation and thus the secretion of several glycoproteins as well as for the synthesis of glycosyl-phosphatidyl-inositol (GPI) anchored proteins.	Proximal digestive tract
EC 5.3.1.8 Mannose-6-phosphate isomerase Isomerases 14	1. Fructose and mannose metabolism 2. Amino sugar and nucleotide sugar metabolism 3. Biosynthesis of secondary metabolites 4. Biosynthesis of antibiotics	Catalyzes the interconversion of fructose-6-phosphate and mannose-6-phosphate and plays a critical role in maintaining the supply of D-mannose derivatives, which are required for most glycosylation reactions.	Pancreas
EC 1.1.1.22 UDP-glucose 6-dehydrogenase Oxidoreductases 14	1. Pentose and glucuronate interconversions 2. Ascorbate and aldarate metabolism 3. Amino sugar and nucleotide sugar metabolism	Catalyzes the formation of UDP-alpha-D-glucuronate, a constituent of complex glycosaminoglycans. Required for the biosynthesis of chondroitin sulfate and heparan sulfate. Required for embryonic development via its role in the biosynthesis of glycosaminoglycans.	Liver & gallbladder
8-amino-7-oxononanoate synthase Transferases 13	1. Biotin metabolism	Catalyzes the decarboxylative condensation of pimeloyl-[acyl-carrier protein] and L-alanine to produce 8-amino-7-oxononanoate (AON), [acyl-carrier protein], and carbon dioxide. It can also use pimeloyl-CoA instead of pimeloyl-ACP as a substrate, but it is believed that pimeloyl-ACP rather than pimeloyl-CoA is the physiological substrate of BioF.	Liver & gallbladder

Table 1: Functions, metabolic pathways and main expression areas related to the APs that presented the greatest impacts. The metabolic pathways were obtained from KEGG, the APs functions were obtained from Uniprot, and the main expression areas were obtained from The Human Protein Atlas.

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.

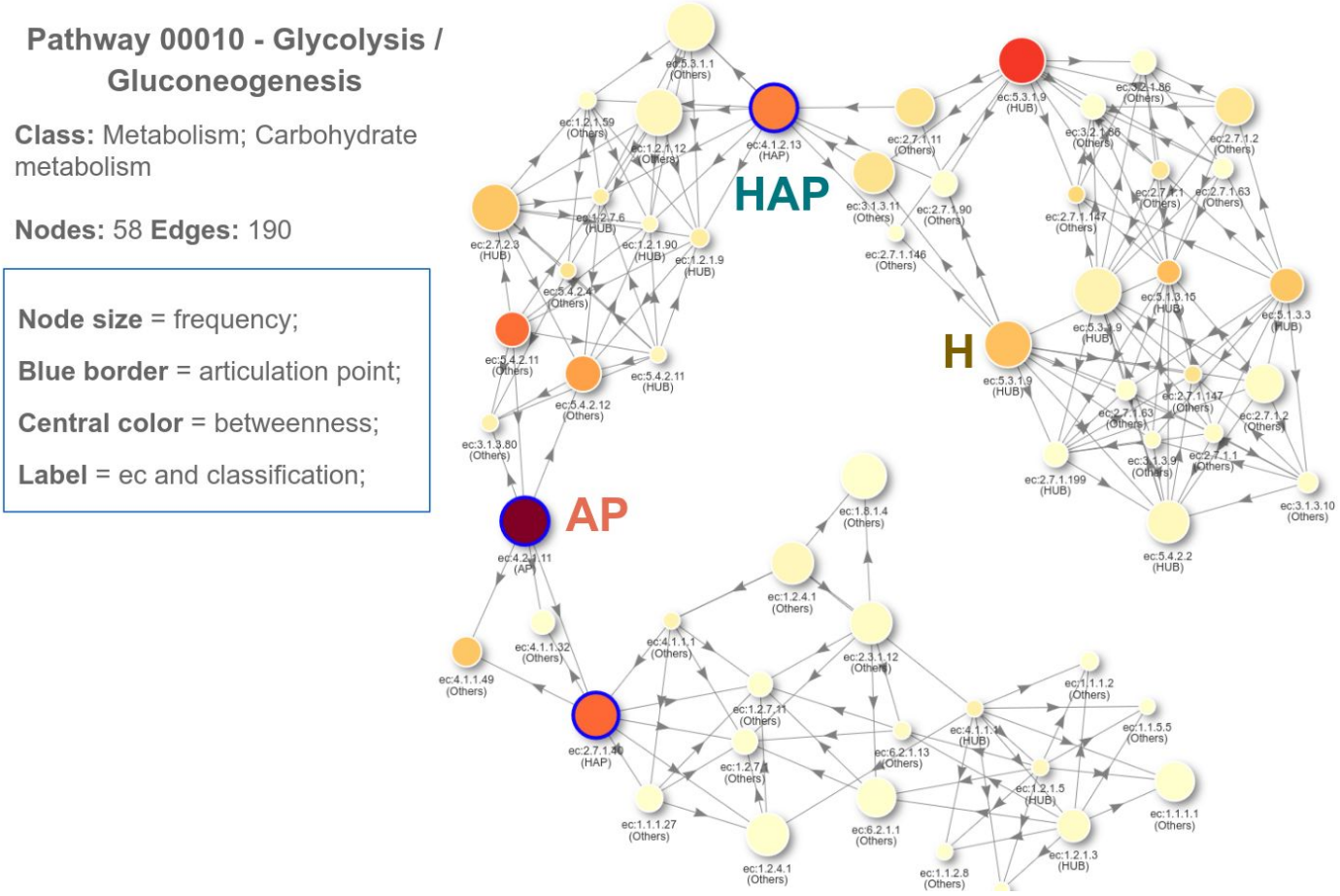


Figure 5: Visualization of pathway 0010 (glycolysis and gluconeogenesis). Nodes represent metabolites and edges represent reactions. Node size represents protein frequency, blue borders indicate APs, color in the center of node indicates betweenness level. The

classification and EC code of the protein are present on the node label. The network was made using 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 on 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 ⅛ 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 S1a) responsible for the defense of some plant species (Supporting material S1b). Also, approximately 31% of metabolic pathways are related to at least ⅝ of KEGG species, this may be due to these pathways being essential in different types of living beings (Supporting material S2), such as glycolysis and gluconeogenesis route 00010, present in 6084 species (Supporting material S3).

Besides, we studied the profile of each protein within the metabolic pathways. HAP represented 7% 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 to each other through a few HAPs. The remotion of one of these HAPs 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 HAPs might be essential proteins. HUB is approximately 17%

of the analyzed proteins. These proteins connect several other proteins of the same community, presenting a high degree. Unlike HAPs, HUBs are not articulation points because a protein cluster has more than one way to connect. As there are usually several protein communities within a network, it is likely that HUBs will be present in larger amounts when compared to HAPs. However, it is important to highlight that the absence of this type of protein does not disconnect the community, but may affect the network in some way.

APs correspond to approximately 13% of the total proteins studied. Although they are important points for the network, normally APs are not as centralized as HAPs. In addition, they do not connect with many neighboring proteins (present low degree). Other proteins correspond to 61% of the proteins because, in terms of network topology, they do not play a crucial role 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 HAPs, such as Benzoxazinoid biosynthesis pathway 00402 (Supporting Information S1), which is composed of only eight proteins, of which six are HAPs. Besides, it is formed by a one-way bridge in terms of mass flow. Other pathways had no HAPs, such as caffeine metabolism pathway 00232 (Supporting Information S4). In this case, there are few communities, which are overcrowded; Some communities are connected to each other through bridges that represent one-way paths. These bridges are composed of APs, which connect with only one protein. Besides, most metabolic pathways are composed of at least three protein classifications. 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 HAPs and APs. This pathway has a considerable amount of APs and HAPs. However, although there are several highly connected nodes in the network central area, none of them an AP or HAP due to the connection handles, which guarantee redundancy in this area of the network.

In the hypergeometric analysis, we observed that there was a considerable concentration of AP in the protein groups with higher frequencies (Figure 3). It is important to highlight that APs are high-frequency proteins, suggesting that these proteins may be more essential to the metabolic pathway when compared with proteins with lower-frequency. In the 1148 proteins with higher frequencies (20.8% of total proteins studied), we found a high concentration of AP (263), suggesting that they were strongly present in most species on different metabolic pathways (Figure 3). The other 80% 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 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, assess 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 explore systematically 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 softwares 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 functional analyses and studies used Gene Ontology to determine the biological process and molecular functions of proteins related to cancer [47, 48]. It is important to highlight that no other study evaluated the main metabolic pathways associated with all APs. Most APs with the greatest impacts are lyases and isomerases (AP classes) expressed in the liver and gallbladder, possibly due to most APs are related to amino acid and sugar metabolism, which occurs in the liver. Besides, monooxygenase (the main AP type) exerts many important functions on the liver, such as a cytochrome P450 activity, inflammatory processes, toxic metabolites metabolism, and drug tolerance [49, 50].

In summary, we suggested that APs have the potential to cause a great impact on biological networks. New studies will be required to establish more relationships between these key proteins (APs) and topological 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 [51]. Another future possibility is the usage of machine learning (ML) techniques 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 HAPs and APs represented more than 20% of the studied proteins. Besides, the hypergeometric findings indicate that most of these APs were placed in a group of proteins with the highest frequencies. Approximately 32% of metabolic pathways are related to less than one thousand species since they are related to a few organisms in the KEGG database. Almost 32% of metabolic pathways are related to at least 5,000 species and these pathways are mainly present into prokaryote (91.4%) due to its variety in KEGG. Most APs are related to amino acid and sugar metabolism. Besides, the monooxygenases are the APs with the highest number of related metabolic pathways. 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.

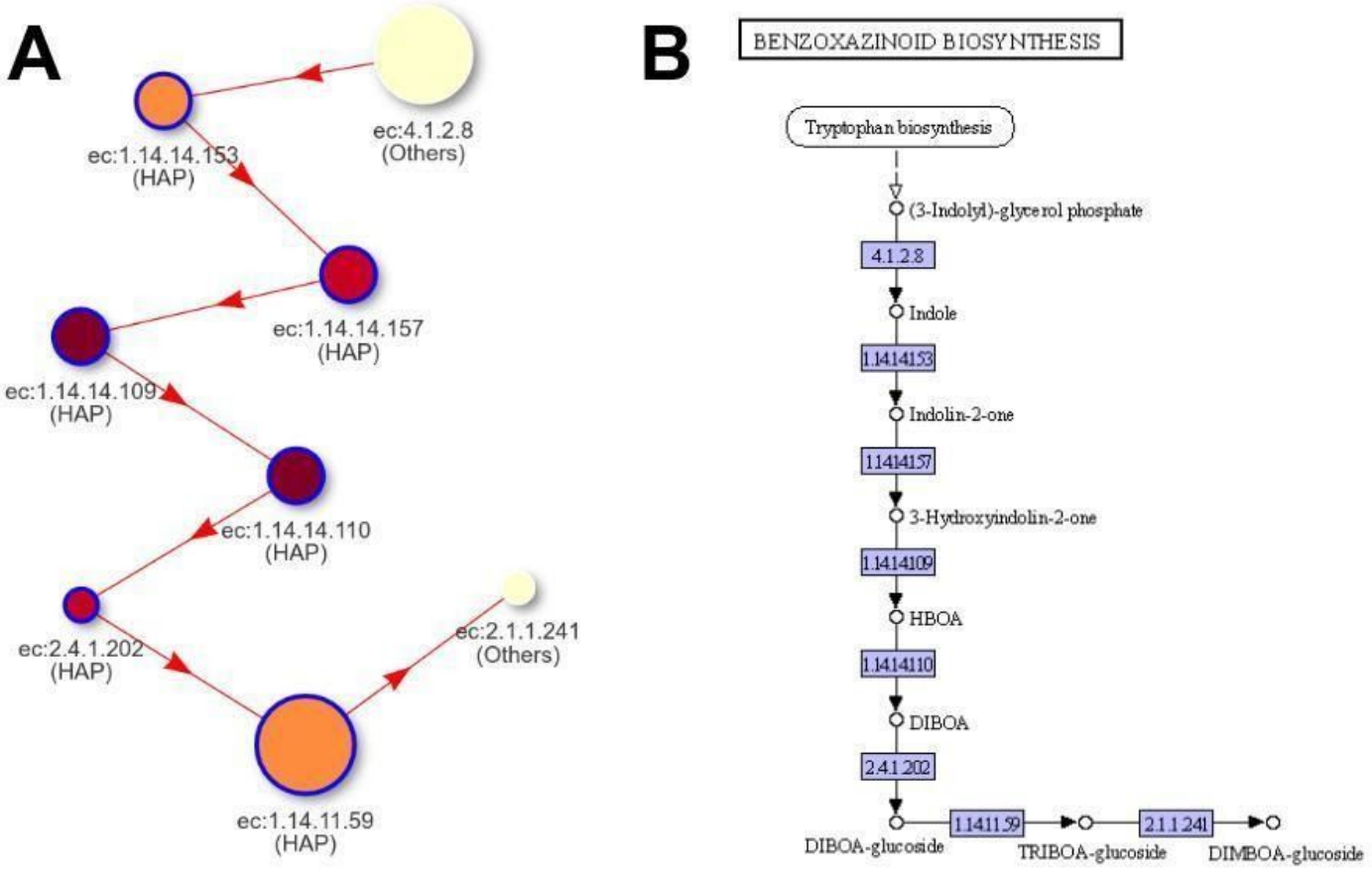
49. Nepomniashchikh VA, Lomivorotov VV, Deryagin MN, Lomivorotov VN, Kniazkova LG. Oxidative stress and monooxygenase liver function in patients with coronary heart disease and multiple organ dysfunction syndrome. European Journal of Anaesthesiology 2009; 26(2): 140-146.

50. Wiśniewska-Knyp JM, Klimczak J, Kolakowski J. Monooxygenase activity and ultrastructural changes of liver in the course of chronic exposure of rats to vinyl chloride. International archives of occupational and environmental health 1980; 46(3): 241-249.

51. 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.

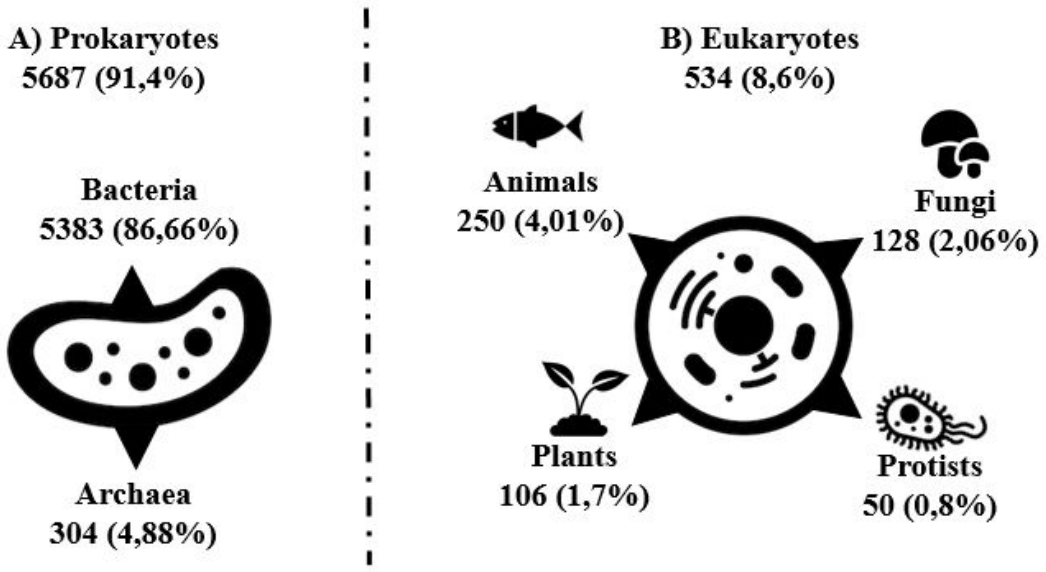
SUPPORTING INFORMATION

Supporting information S1: Pathway 00402 - Benzoxazinoid biosynthesis*

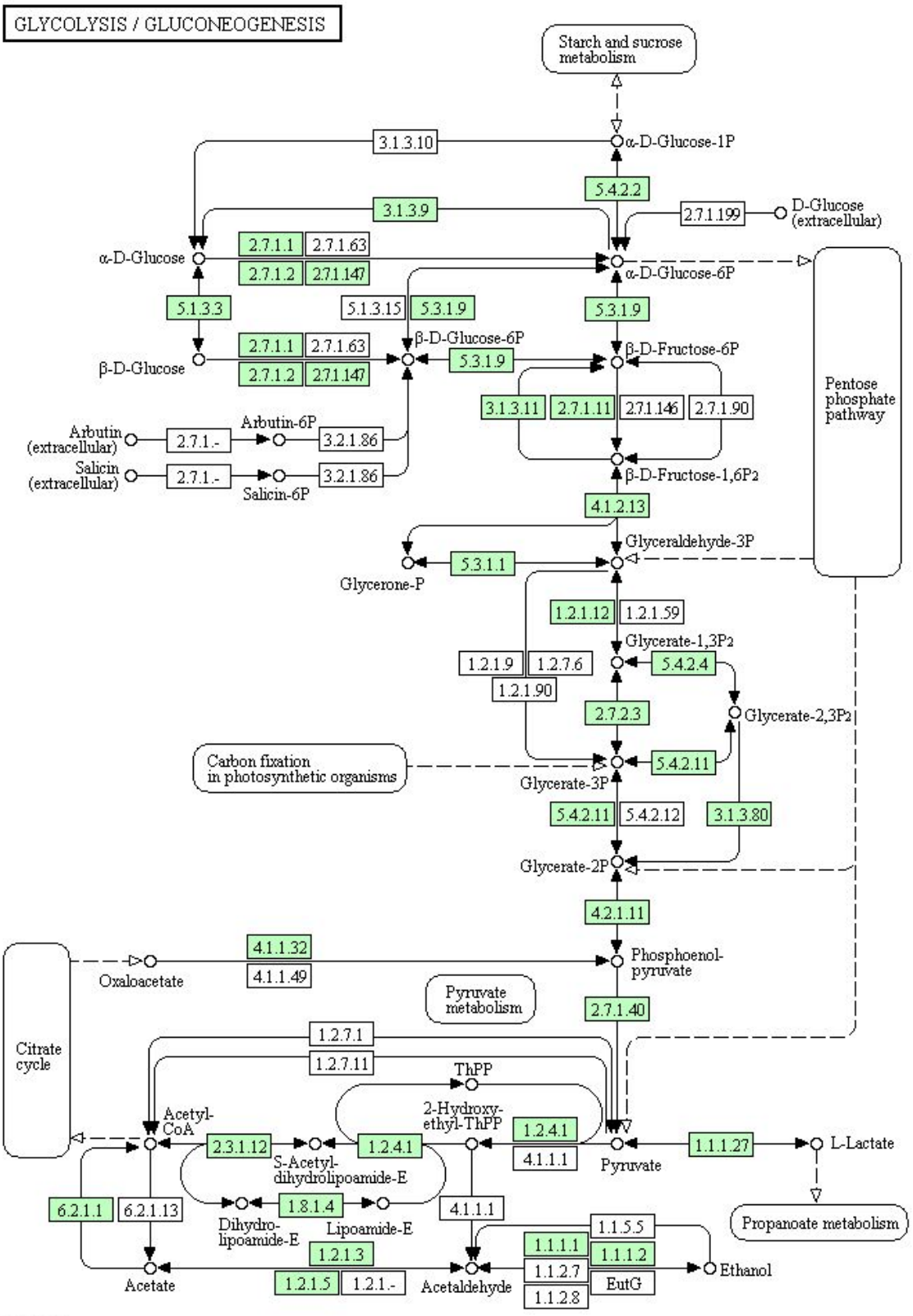


***Nodes:** 8 / **Edges:** 7

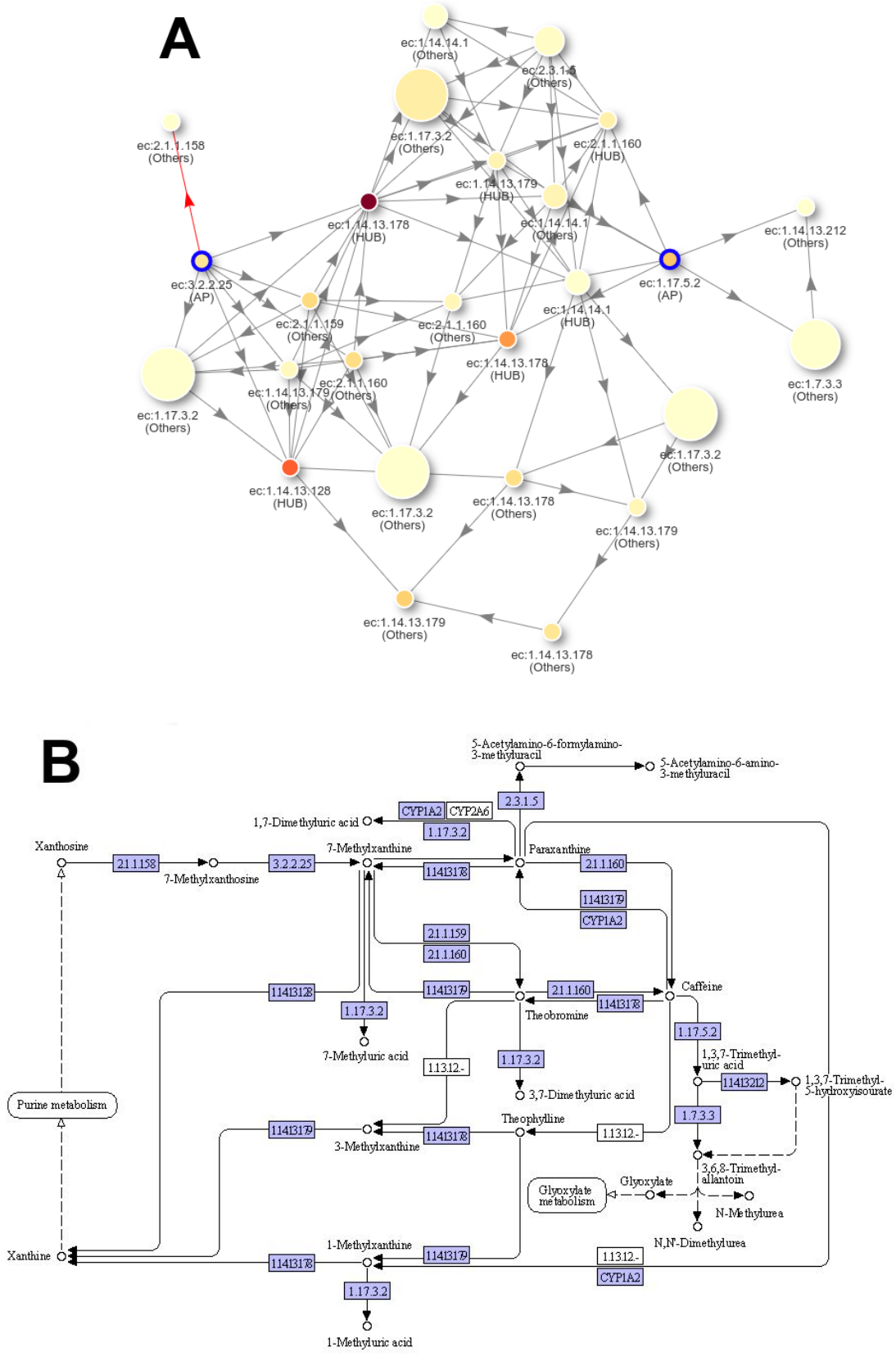
Supporting information S2: Classification and quantity of organisms used in the study. (A) Prokaryotes; (B) Eukaryotes.



Supporting information S3: 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.

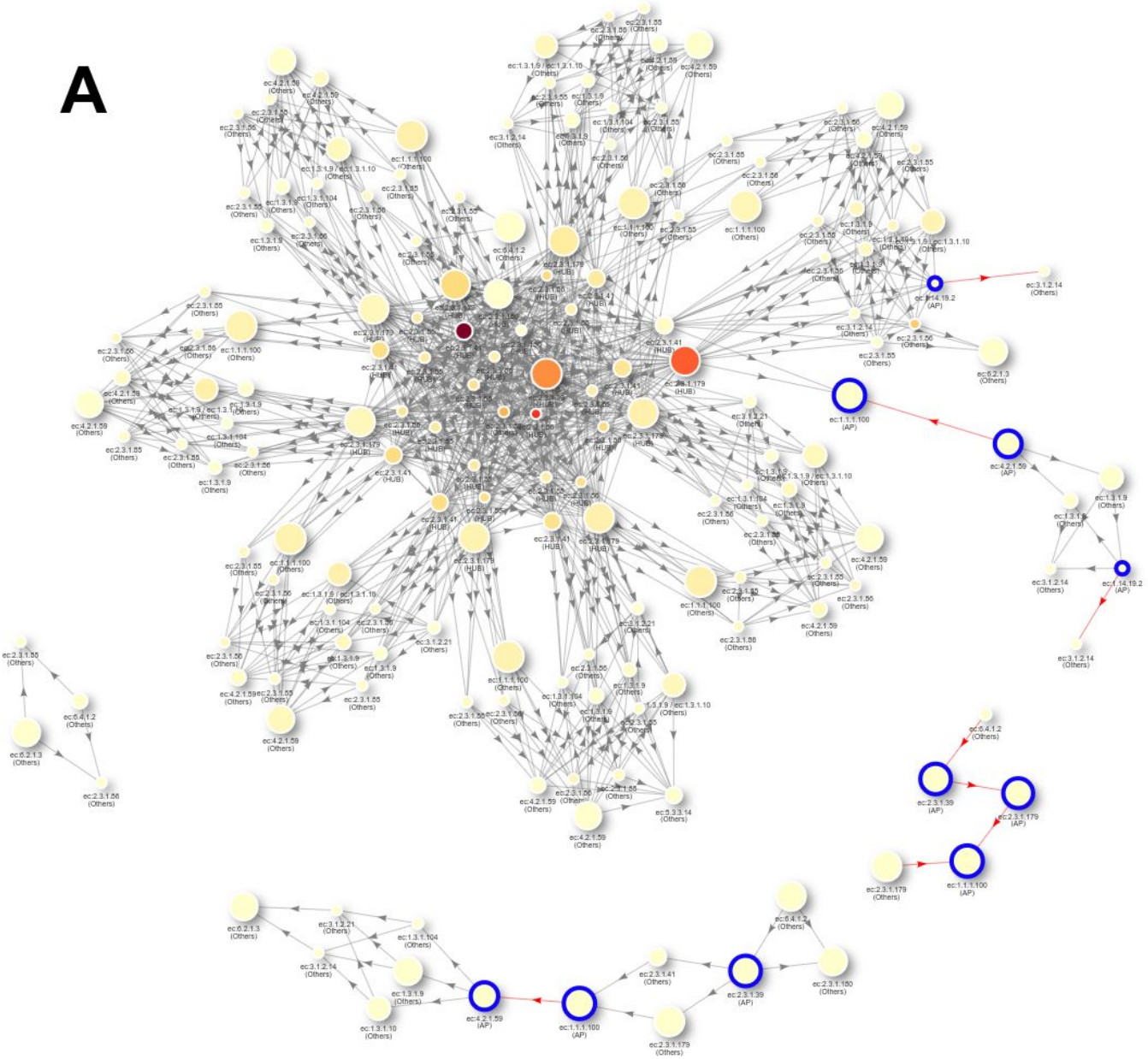

GLYCOLYSIS / GLUCONEOGENESIS
Starch and sucrose metabolism
α-D-Glucose-1P
3.1.3.10
5.4.2.2
3.1.3.9
2.7.1.199
D-Glucose (extracellular)
α-D-Glucose
2.7.1.1
2.7.1.63
2.7.1.2
2.7.1.147
α-D-Glucose-6P
5.1.3.3
5.1.3.15
5.3.1.9
5.3.1.9
β-D-Glucose-6P
β-D-Glucose
β-D-Fructose-6P
Pentose phosphate pathway
3.1.3.11
2.7.1.11
2.7.1.146
2.7.1.90
Arbutin (extracellular)
2.7.1.-
Arbutin-6P
3.2.1.86
Salicin (extracellular)
Salicin-6P
β-D-Fructose-1,6P2
4.1.2.13
Glyceraldehyde-3P
Glycerone-P
1.2.1.12
1.2.1.59
Glycerate-1,3P2
1.2.1.9
1.2.7.6
1.2.1.90
5.4.2.4
2.7.2.3
Glycerate-2,3P2
Carbon fixation in photosynthetic organisms
5.4.2.11
Glycerate-3P
5.4.2.11
5.4.2.12
3.1.3.80
Glycerate-2P
4.2.1.11
4.1.1.32
4.1.1.49
Oxaloacetate
Phosphoenol-pyruvate
Pyruvate metabolism
2.7.1.40
Citrate cycle
1.2.7.1
1.2.7.11
ThPP
Acetyl-CoA
2-Hydroxy-ethyl-ThPP
1.2.4.1
2.3.1.12
1.2.4.1
4.1.1.1
Pyruvate
1.1.1.27
L-Lactate
S-Acetyl-dihydrolipoamide-E
6.2.1.1
6.2.1.13
1.8.1.4
4.1.1.1
Dihydro-lipoamide-E
Lipoamide-E
Propanoate metabolism
1.1.1.1
1.1.1.2
1.1.5.5
1.2.1.3
1.1.2.7
EutG
Ethanol
Acetate
1.2.1.5
1.2.1.-
Acetaldehyde
1.1.2.8

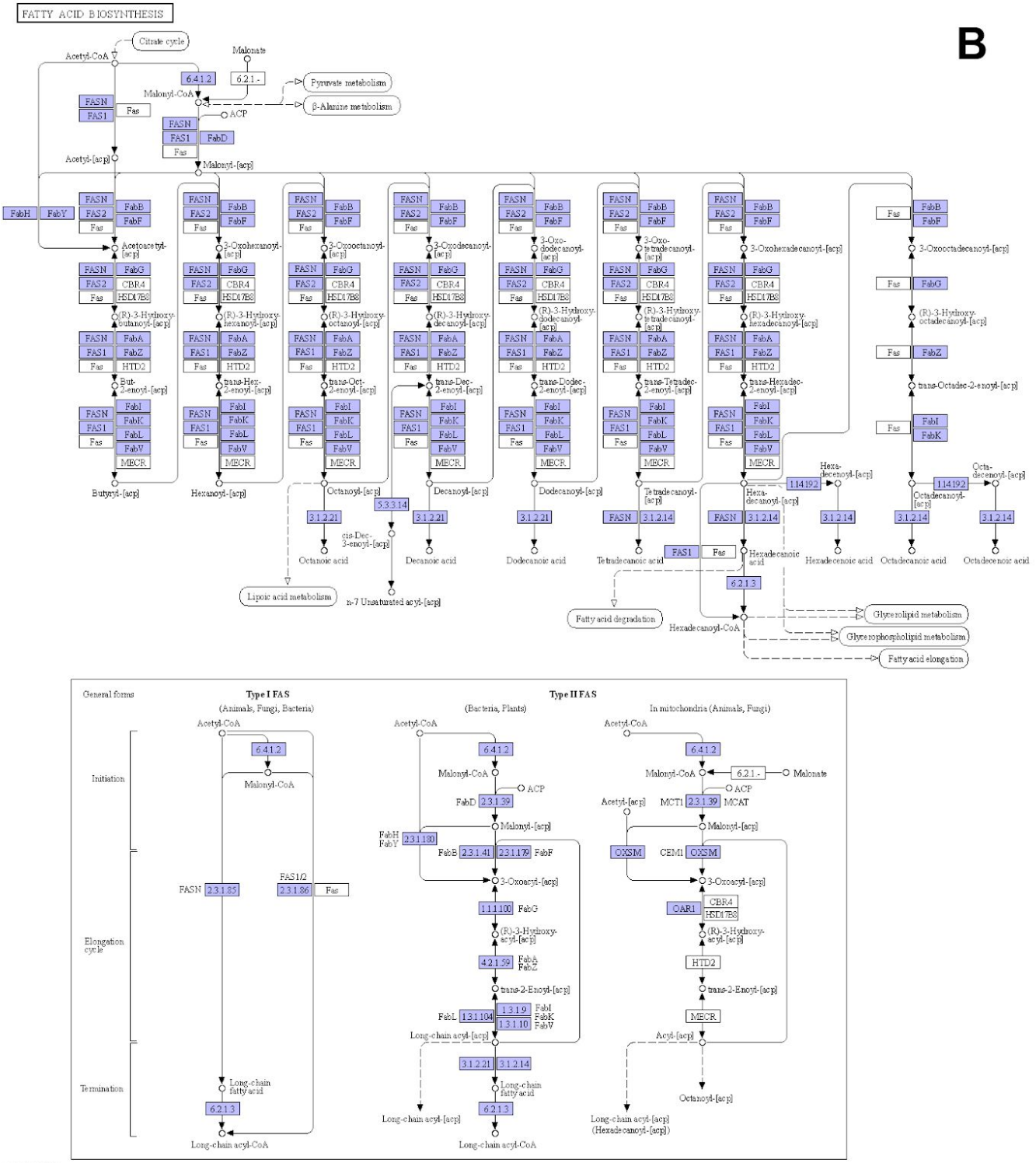
Supporting information S4: Pathway 00232 - Caffeine metabolism*



***Nodes:** 26 / **Edges:** 76

Supporting information S5: Pathway 00061 - Fatty acid biosynthesis*





***Nodes:** 170 / **Edges:** 1063