Functional & Integrative Genomics (2022) 22:1433–1448
https://doi.org/10.1007/s10142-022-00907-y

ORIGINAL ARTICLE

Basics on network theory to analyze biological systems: a hands-on outlook

Gerardo Ruiz Amores[1] · **Agustino Martínez-Antonio**[1]

Received: 21 January 2022 / Revised: 3 October 2022 / Accepted: 4 October 2022 / Published online: 13 October 2022

Abstract

Biological processes result from interactions among molecules and cell-to-cell communications. In the last 50 years, network theory has empowered advances in understanding molecular networks' structure and dynamics that regulate biological systems. Adopting a network data analysis point of view at more laboratories might enrich their research capacity to generate forward working hypotheses. This work briefly describes network theory origins and provides basic graph analysis principles in biological systems, specific centrality measurements, and the main models for network structures. Also, we describe a workflow employing user-friendly free platforms to process, construct, and analyze transcriptome data from a network perspective. With this assay, we expect to encourage the implementation of network theory analysis on biological data in everyday laboratory research.

Keywords Network construction · Network analysis · Galaxy · Cytoscape · Transcriptome data

Introduction

Biological processes such as movement, development, disease transmission, breathing, and thinking rely on molecular, cell-to-cell, or signaling communications. Scientists have understood those mechanisms through biochemistry, molecular, and cellular biology studies. Advances in identifying metabolic, signaling pathways, and proteins participating in those mechanisms contributed to a partial picture of phenomena. This reductionist approach was necessary to understand the function of individual elements and realize that some of them have multiple or redundant roles in biological systems.

To mention some examples, the biotechnological synthesis of L-tryptophan (TRP) has been a focus of attention for many years. Tryptophan is an amino acid of high biological and commercial value due to its essentiality in biological systems. It is a precursor of developmental hormones or additives in the food industry. The continuous genetic improvements have not fully exploited the bacterial metabolic capabilities to produce this amino acid maximally. This circumstance is due to the multiple metabolic and genetic crosstalk that mediates its production (Crawford 1975). Another example relies on sex plasticity in some organisms, such as fishes, flies, or wasps. Usually, sex is determined before birth; however, some species can switch their sexual phenotype depending on external inputs such as temperature or food availability. Understanding such plasticity is essential for industrial or biotechnological applications. Sex determination is a process of high complexity due to multiple genes and their interactions involved. Lastly, the decrease or absence of the P53 protein is a usual hallmark of cancer. However, its molecular role in numerous biological processes such as cell cycle, genetic stability, and apoptosis makes studying and therapy based on the P53 complex (Cheah and Looi 2001). Therefore, there is a growing necessity to implement alternatives that allow understanding the structure and dynamics of the molecular networks related to the above described and other biological processes.

With the advance of new technologies such as the genome, metabolome, and proteome analysis, there is a need to implement alternatives to analyze the bulk of produced data (Alm and Arkin 2003). It has been helpful to adopt some physics principles of analysis, such as network theory,

✉ Agustino Martínez-Antonio
agustino.martinez@cinvestav.mx

[1] Biological Engineering Laboratory, Genetic Engineering Department, Center for Research and Advanced Studies of the National Polytechnic Institute (Cinvestav), Campus Irapuato, Km. 9.6 Libramiento Norte Carretera Irapuato-León 36821, Irapuato, Guanajuato, México

to explore massive data and complex systems. A complex system is constituted of a set of units with individual characteristics which respond and adapt to various conditions providing robustness to the system. However, the overall system response is difficult to predetermine by studying their components (Amaral and Ottino 2004). Therefore, from a holistic perspective, we analyze complex systems with network theory by representing the biological units as nodes and their interactions as links, generating a graph (Fig. 1). This representation permits studying and analyzing organisms' functional structure through different models. With these analyses, we can identify essential characteristics such as the most connected units in the system, the route most information disseminates or linked units as modules that perform a particular function (Barabási and Oltvai 2004). This abstraction has also allowed analyzing economic, logistic, social, internet, and ecological phenomena. Researchers have applied network theory in biology since 50 years ago to extend the understanding of disease transmissions and diffusion, proposing new functions for unknown proteins or identifying genes with implications in diseases (Yu et al. 2013). This assay provides background on network theory. The detailed mathematical foundation is found in the literature cited. Hence, we will describe some principles and metrics to explain how network theory could study biological systems. Following this, we provide a hands-on tutorial on user-friendly platforms to process, construct, and explore a pair of natural phenomena employing transcriptomic data and network theory. We warm to differentiated network theory analysis from other approximations, such as the genomic reconstruction of metabolic networks in which biochemical reactions are transformed into a 0, 1 matrix to calculate and optimize physiological conditions (Palsson 2015).

A brief history of graph theory

The beginning of the graph theory is blurred in time. The first graphical representations using nodes and links date from the first century BC. Marcus Tullius Cicero was one of Rome's greatest orators. To achieve an outstanding speech quality, he recommended abstracting parts of the discourse into figures and putting those as physical components. When necessary, he provided a direction among these components into a stage. The result was an ordered and fluent lecture (Fig. 2A) (Barnes and Harary 1983). This memorizing method became a mnemonic technique to improve memory and possibly the first graph representation by abstracting the speech pieces into figures' "nodes" and their structural logic into "links" with a direction.

In 1736, the mathematician Euler published his works describing Königsberg's bridge paradigm (Zaytsev 2008). He explains that in Russia, the Kneiphof island comprising the Königsberg city was isolated because of two rivers. There were seven bridges to get in or out of the town (Fig. 2B). Then, he wondered whether it would be possible that a person could establish a route to cross each of the bridges only once. To solve it, Euler abstracted the geographic zones to nodes, and the bridges to access the island were represented with lines (Fig. 2C). Then he classified the nodes into odd or even, based on the residue, dividing the number of links by 2. Then, whether the result was an

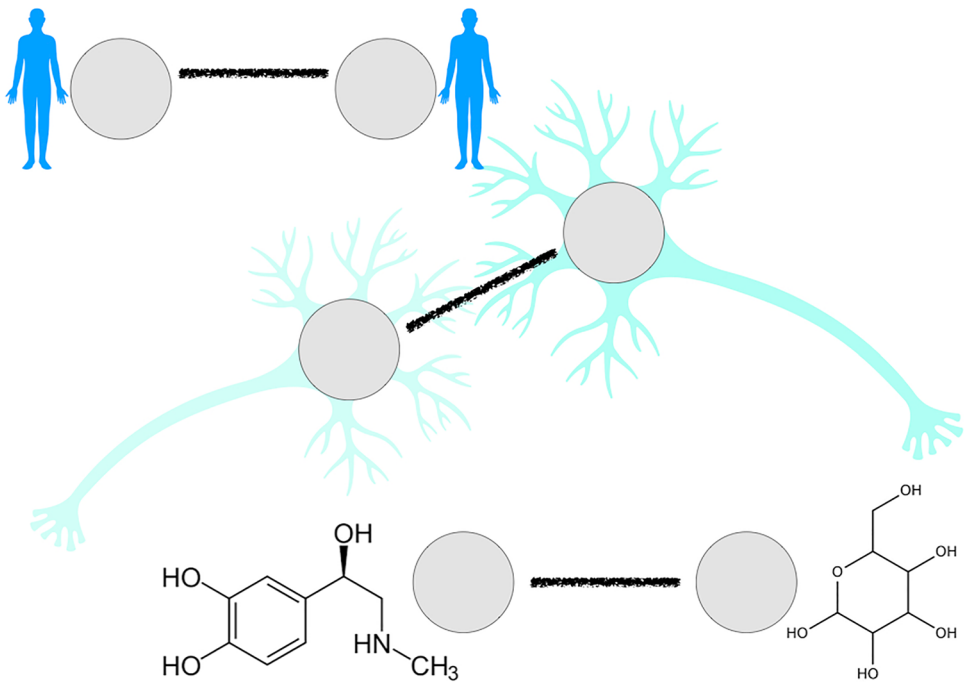

Fig. 1 Biological phenomena are abstracted into a graph. The circles represent nodes or N (any physical unit can be represented in this way, such as humans, neurons, or molecules). Any interaction, such as friendship, metabolic reaction, biological, or chemical interaction, can be represented by a line joining two nodes called a link or edge or K (shaded black bars)

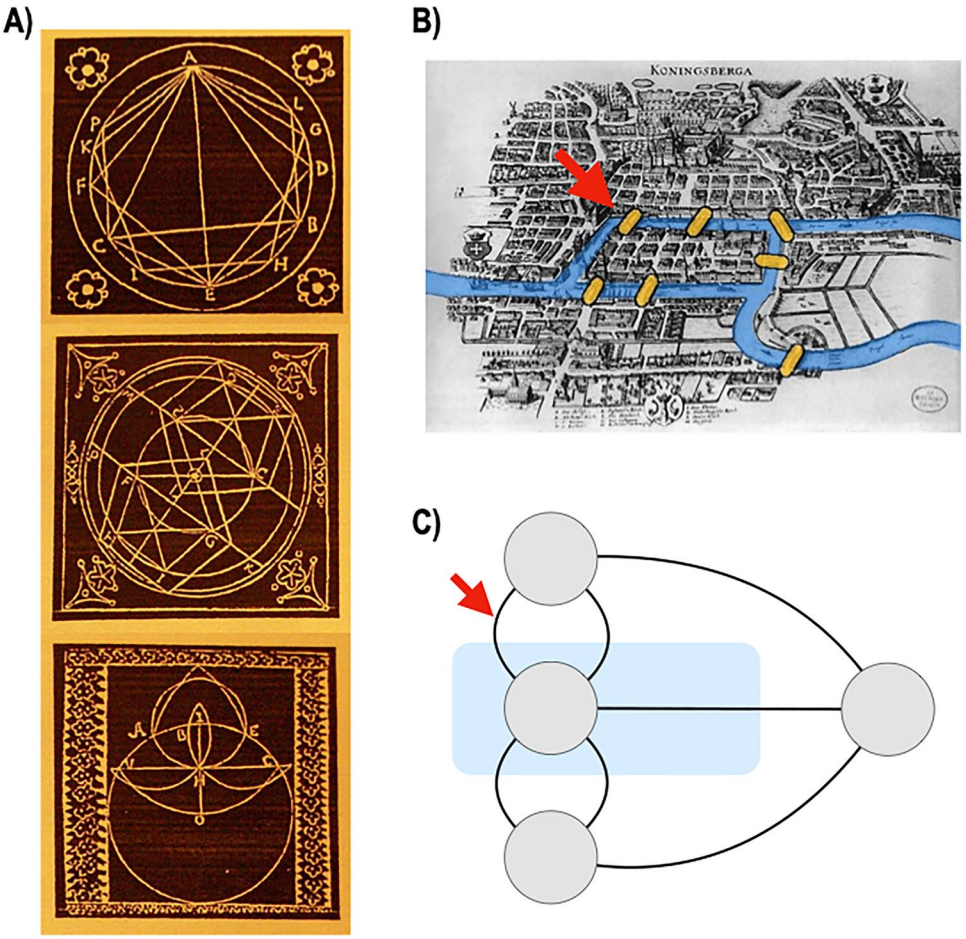

Fig. 2 Graph theory origins. **A** Work of cosmologist Giordano Bruno on the mnemonic method (org/wiki/File: Memory-seals). **B** Königsberg's city and the bridge paradigm. **C** Graph theory was implemented by the mathematician Euler in 1976; he abstracted the landscape into nodes and bridges into lines. The blue square represents the island. The red arrow shows the bridge abstraction into a line (from **B** to **C**); the same applies to the remaining six bridges

integer, it was possible a path (just one walkthrough); fractions were not. Euler demonstrates that a graph can have specific features such as just two kinds of nodes based on its connections, integer and fractional (Amaral and Ottino 2004). Later, in 1850, graph theory helped describe the structure of electronic circuits and chemical isomerism (Barnes and Harary 1983; Estrada 2013). In 1950, Kochen and Pool proposed applying graph theory to human societies to understand its structure. People were abstracted to nodes and links as a possibility to establish a relationship between any pair of nodes, creating randomized networks (de Sola Pool and Kochen 1978). In 1969, Erdos and Rényi defined a normal distribution of connections, supporting the "small world" effect that explains the six-degree separation phenomenon (Erdos and Rényi 1960; Milgram 1967). Finally, in 1999, Barabási and Albert enunciated a new graph in social networks, called the "scale-free." This model evidence the existence of many nodes with few links and rare with many connections. This distribution is also known as the power law and is demonstrated in biological systems also (Barabási and Oltvai 2004).

Types of biological graph networks—nomenclature

Abstraction of the units of an organism to nodes makes it possible to apply network theory to analyze different phenomena such as metabolic fluxes, protein–protein interactions, and gene transcription, among others. Among the different types of graphics, the principal ones found in biological systems are undirected, directed, and weighted (Fig. 3) (Barabási and Oltvai 2004).

In the first case, undirected graphs, a couple of nodes are linked with a line without direction and is the most basic network representation. Here, there is no information regarding interactions among the nodes in the graph (i.e., activation or repression). Still, it is known that a relationship exists between these nodes (e.g., steric effect). In that case, a link connecting two nodes is the best choice to represent such interaction. An example is a protein–protein interaction between scaffold proteins in a signaling pathway or a protein interchange with small molecules such as cofactors (Fig. 3A, D).

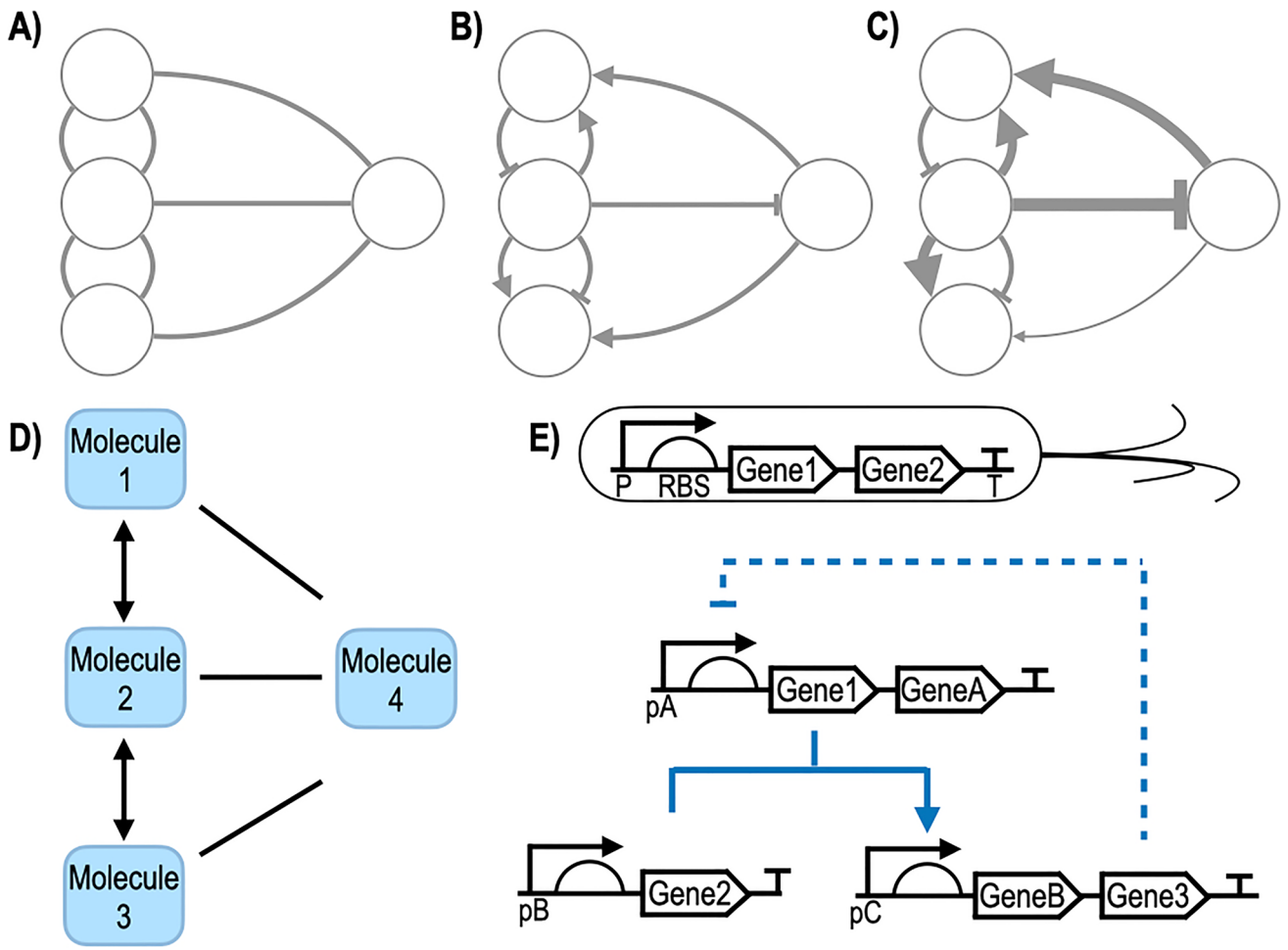

Fig. 3 Standard biological graph networks. **A** Undirected, **B** directed, and **C** directed-weighted graphs; a line represents a link between two nodes, directed arrows represent activation, and those for repression show a crossline at the end. The gross of arrows in (**C**) indicates the relative link weight. **D** Scaffold proteins, small molecules, or others with no complete knowledge of the kind of interaction (lines) as in the case of Molecule 4 in a pathway are better described using an undirected graph. **E** A bacterial genetic circuit; P is a promoter genetic region; RBS is ribosome binding site. Genes followed by numbers represent to transcription factors, and those followed by letters denote any other gene product, such as protein or ribonucleic acid. T is a translation terminator sequence. Here, the products of Gene 1 and Gene 2 induce the production of Gene B and Gene 3; this last will eventually turn off the Gene 1 and Gene A operon

With the work of researchers, the output of many chemical reactions or interactions among proteins, nucleic acids, or protein-nucleic acids is identifiable. Such biological interchanges include phosphorylation, gene regulation, and enzymatic reactions, to mention some. This precise knowledge makes it possible to assign a direction (i.e., negative or positive) to such interaction to build a directed network (Fig. 3B). For instance, transcription factors (TF) commonly regulate gene expression in bacteria since TF recruits or interferes with the RNA polymerase, inducing activation or repression (Fig. 3E). In this context, different biochemical and molecular techniques, such as footprinting or promoter activity assay, help to confirm those regulatory interactions (Amores et al. 2017). The outcome can commonly be displayed in the graph with an arrow for activation or a crossline at the end for repression (Fig. 3B, E). It should be advised not to mix interactions representing different biological processes in the same network, mainly those accomplished at different time scales, to be analyzed with graph theory tools.

Finally, the weighted graphs examine those in which the link between two nodes has a relative value (Fig. 3C, E). A link weight can be given from protein–protein interactions using nucleic magnetic resonance (NMR) or X-ray crystallography to establish weak or strong relationships among proteins or chemical functional groups. For gene expression-weighted networks, the importance of the link between a couple of nodes can be inferred massively utilizing the transcriptomic data from next-generation sequencing (NGS) or in TFs by the number of regulated genes (Fig. 3C, E).

When relations are deduced for the first time, the process is named "de-novo," as explained below in the model section ("Topological properties of biological networks—molecular and cellular models" section).

Centrality measurements in biological networks—graph properties

Once a network is built, it is possible to analyze it based on the intrinsic properties of the graph. For instance, nodes' connectivity and weights are used to determine the inherent graph structure upon network construction using NGS data (or related methodologies). Those are known as centrality measurements and are studied using different algorithms based on connectivity properties. The most common measures are degree, betweenness, closeness, and clustering (Fig. 4). Then, centrality measurements allow for identifying distinctive nodes with specific connectivity properties.

The degree measure indicates the number of connections in a node (Fig. 4A). It is helpful to identify the most and least connected nodes; its distribution distinguishes between different classes of networks (see "Topological properties of biological networks—molecular and cellular models" section). This distribution is possible by estimating each node's links in the entire network. Then a histogram is created with those data being helpful to distinguish between network models to which a network best fits (Fig. 5). The in-degree analysis denotes the number of links that point to a node, highlighting the nodes which receive more incoming connections (Fig. 4B). Meanwhile, the out-degree indicates the number of links that start from a node, featuring the nodes that emit more connections (Fig. 4C) (Barabási and Oltvai 2004).

Betweenness centrality is a measurement to estimate how critical a node is as a bridge of information over the rest of the network. It measures the proportion of nodes in the network over which a node has influence compared to the rest (Estrada 2013). Also, this centrality measure indicates which nodes spread quickly or control the information flux better. Then, blocking or removing these nodes will disrupt communications in the network more effectively (Fig. 4D).

In link steps, the closeness centrality measures the average a node is to the rest of the nodes in the network. A high

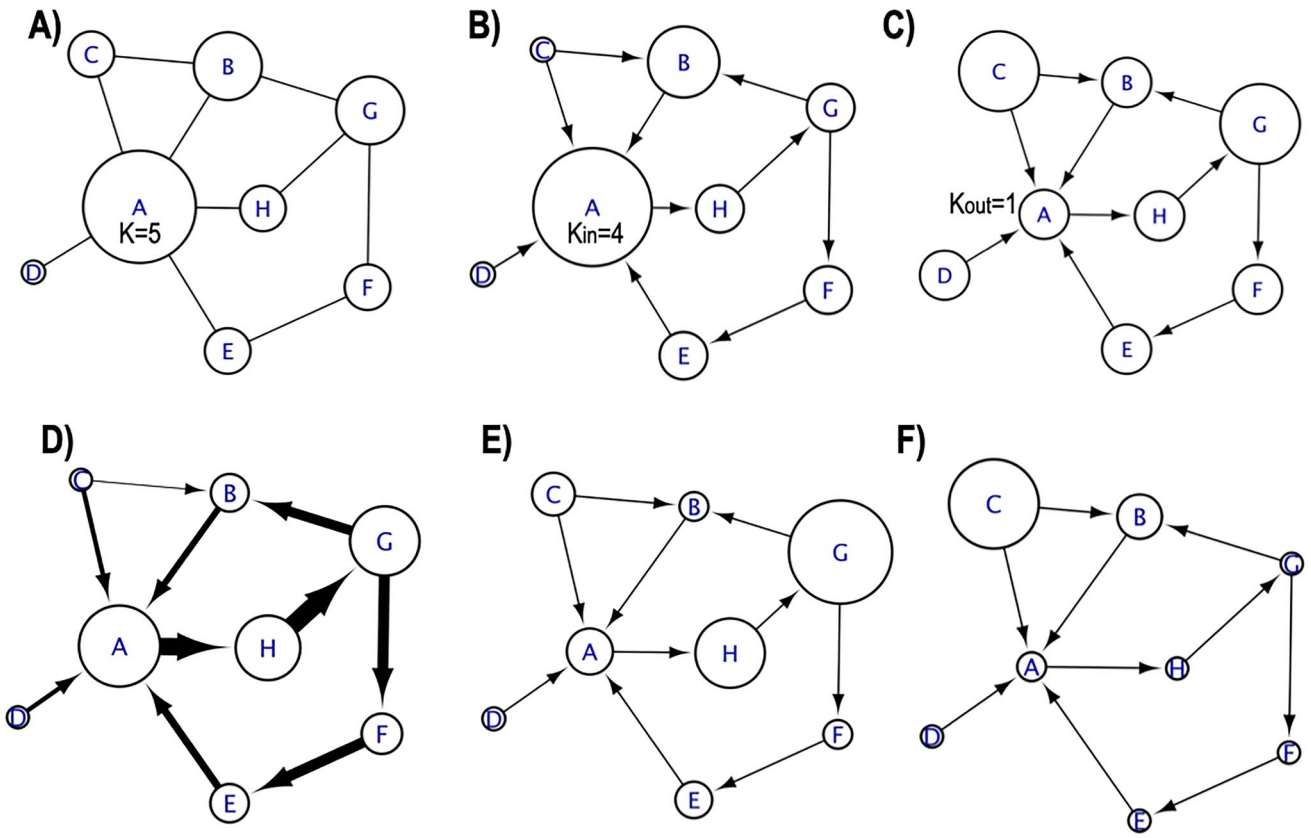

Fig. 4 Graph centrality measurements. The data for the graphs was taken from Barabasi and Oltvai (2004) and loaded into the cytoscape.org software. Then we performed a centrality analysis, and the resulting maps are shown for **A** degree, **B** in-degree, **C** out-degree, **D** betweenness, **E** closeness, and **F** clustering coefficients (note as the relevance of the nodes change in each centrality measure). K indicates links/edges

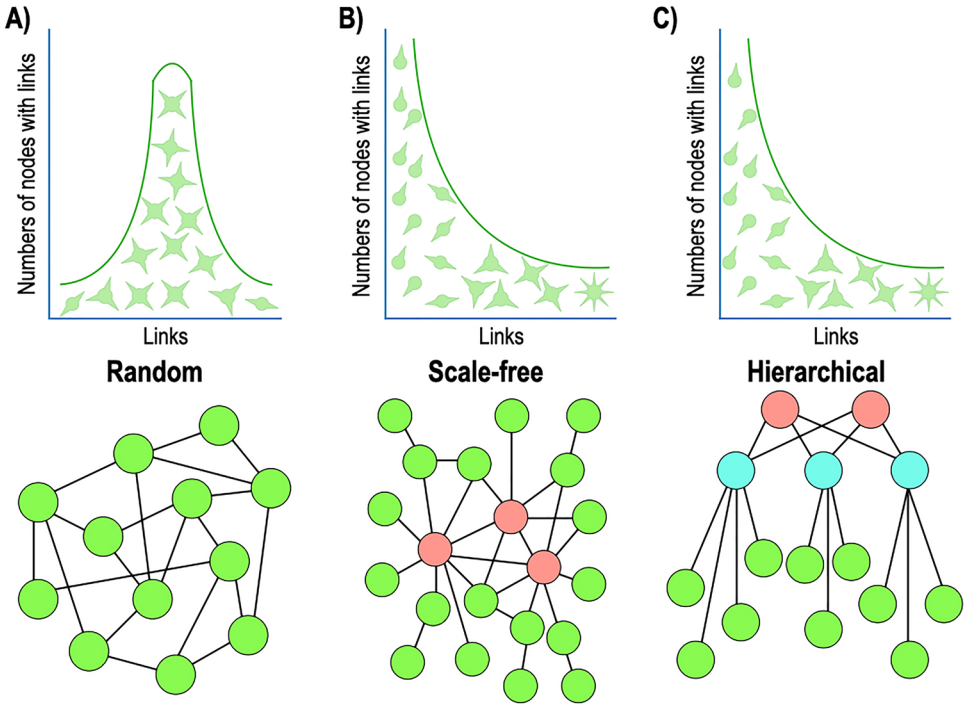

Fig. 5 Biological graph models. **A** Random networks (Erdos and Rényi 1960), **B** scale-free networks (Barabási and Albert 1999), and **C** hierarchical networks representations were gathered from (Barabási and Oltvai 2004; Barabási 2002; Ravasz and Barabási 2003). At the top is the Poisson and power-law distribution of nodes with connections representing the differences between **A** random, **B** scale-free, and **C** hierarchical networks. At the bottom is the topological representation of a network with (**A**) where nodes have a similar number of connections, **B** distinctive hubs due to more connections in a Scale-free network, and **C** hierarchical structure in a network showing modules or groups of nodes. Data of networks were taken from Barabasi and Oltvai (2004)

value represents the most distances to all other nodes and vice versa (Fig. 4E).

The clustering coefficient is the ratio of the number of triangles incident to a node respecting the maximum possible number of such triangles in the entire network (Fig. 4F) (Estrada 2013). In other words, this coefficient determines nodes that tie together to perform similar tasks, for example, genes forming feed-forward loops networks' motifs (Itzkovitz and Alon 2005).

The shortest path represents the path with the minimum number of edges between any two nodes in a network. Its mean determines the number of steps along the shortest paths in a network. It could be helpful to select a route to spread information from one node to another more quickly or the minimal enzymes/genes required in metabolic pathways.

Topological properties of biological networks—molecular and cellular models

Network models represent the main properties observed in complex systems. These are random, scale-free, and hierarchical networks (Fig. 5).

In the random networks (Erdös and Rényi 1960), the degree distribution of nodes follows a Poisson distribution (Fig. 5A). Therefore, this network will present most nodes with intermediate numbers of connections and decreasing nodes with fewer or most links. This model is more typical in network engineering but hardly explains the topological architecture of networks inside an organism.

In the scale-free network, the degree distribution among nodes follows a power-law distribution, implying the existence of nodes with different order magnitudes of connections (Fig. 5B). In this kind of graph, many genes or proteins have few links and sporadic nodes with a more significant number of interactions. Most biological networks, such as metabolic, genetic, regulatory, and protein–protein, are scale-free.

Therefore, it is possible to observe hierarchical genetic regulatory networks as in *Escherichia coli*, where some transcription factors interact with many genes; these are called global regulators (Fig. 5C) (Martínez-Antonio and Collado-Vides 2003).

Hands-on network analysis, a user-friendly workflow approach

Of the different high-throughput techniques, transcriptome analysis from NGS data is probably the most demanded and challenging (Bohra et al. 2021). Their research requires bioinformatic abilities and deep knowledge of the studied phenomena. This technique allows for identifying over or down-expressed genes in an organism, providing information for specific genes or a group of genes with similar expression patterns, suggesting a function in the observed biological phenomena. Because of this importance, we design a workflow based on the Galaxy platform (https://usegalaxy.org/) and Cytoscape (https://cytoscape.org/) graphical user interface (GUI) to process, construct, and analyze transcriptomic

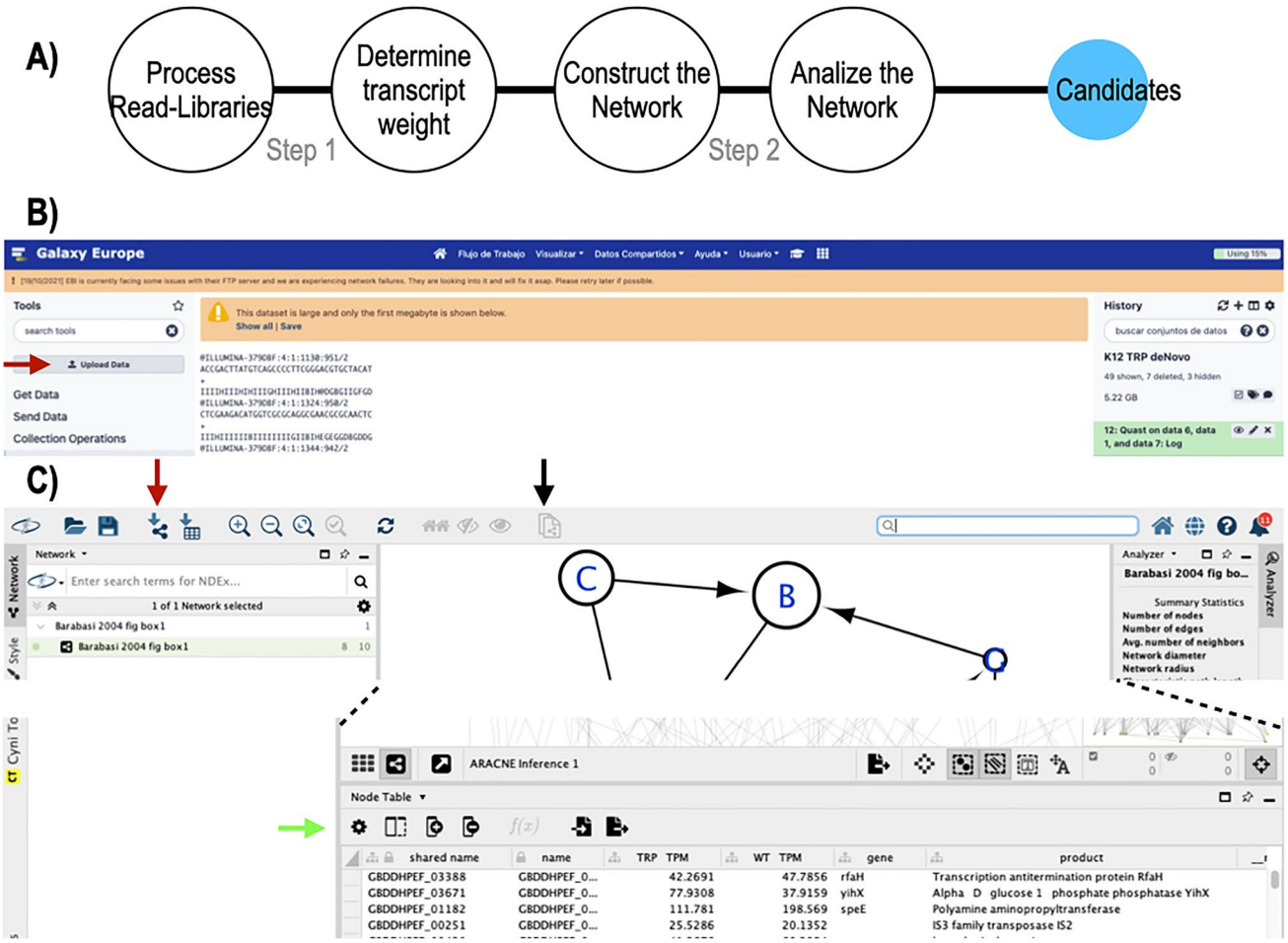

Fig. 6 Network workflow and user-friendly platforms. **A** A pipeline to process, construct and analyze NGS data using Network theory described in the main text. **B** Galaxy web platform, we use the red arrow indicating the upload tool, upper the applications tool searchbox. **C** Cytoscape software graphical interface is shown where the red arrow shows the upload network option. The black arrow shows the extraction nodes and edges tool. The green arrow shows the node/edge/network table graphical option

data from a network perspective. This workflow consists of two steps (Fig. 6A). The first phase executes at Galaxy to get quality reads and estimate expression value on annotated transcripts (Taylor et al. 2007). The second step merges the quantification values for each gene at the different experimental conditions, and the network is constructed and analyzed in Cytoscape (Shannon et al. 2003).

Therefore, to provide a hands-on experience, we will process, construct, and analyze transcriptome data from two different organisms. First, we will consider the bacterium *Escherichia coli* when growing in the absence or presence of tryptophan (Bordbar et al. 2014), aiming to identify the structure of the genetic network to produce this valuable metabolite. Second, we interrogate the transcriptome of the parasitoid *Diachasmimorpha longicaudata*. This wasp infects a fruit-fly pest, employing it as a biological control agent. This parasitoid has complementary sex determination and multiple sex loci mechanisms, resulting in male-favored populations. Identifying genes related to the sex-determination system might improve the massive rearing protocols for female organisms (Carabajal Paladino et al. 2015). Throughout network construction, a key feature is knowing the methodologies and sequencing platforms for selecting the proper analyses' algorithms. First, we must choose the appropriate library preparation and platform technology and the research aim in our biological phenomena. Once having these in mind, we can proceed with the analysis.

Process to read libraries at Galaxy

The genetic material extracted from any biological sample, usually designed as control and experimental conditions, is DNA sequenced. The raw data, corresponding to randomized fragments of the nucleotide sequences of interest, commonly carries out additional genetic elements necessary to the technique, such as adaptors and primers. Therefore,

a trimming step of undesired sequences is essential. It is possible to get raw data from different repository databases, such as the National Center for Biotechnology Information (NCBI) (Sayers et al. 2021).

The Galaxy platform is an open-access web-based platform for performing the accessible, reproducible, and transparent genomic analysis (Taylor et al. 2007; Goecks et al. 2010). Although we can work without a login, we recommend creating a free account. It gives us more advantages, like more memory space to run our workflow. This web platform is divided into three sections (Fig. 6B). Multiple tools permit bioinformatic analysis in the left panel, ranging from single to multiple genes. At the center, the board allows for visualization of any selected tool's customizable options or the database product of the performed analysis. At the right are the logic processes of the workflow and from where we can download the results (Fig. 6B).

First, we should upload the data. There are different options to do it; there are tutorials to explore the various possibilities on the platform. We will assume that we are working with transcriptomic data obtained from our experimentation. Then, we should have the files in FASTQ format on our computer. We recommend keeping independent records for *E. coli* or *D. longicaudata* processes and assigning their respective names. We start by clicking on the "Upload Data" option at the left of the platform. A dialogue window will appear at the center to search into local files to read the library and "start" the upload. Here, we download at our computer the FASTQ libraries from NCBI of *Escherichia coli* K-12 grown on minimal media in the absence and presence of tryptophan (SRR922261 and SRR922264, respectively) (Bordbar et al. 2014). For male larvae, female, and male developmental stages of *D. longicaudata*, it corresponds to SRR3336273.1, SRR3336336.1, and SRR3336337.1, of the contigs constructed in the Bio project PRJNA317427-GELG01.1 (Mannino et al. 2016).

As stated above, known methodology and platform are fundamental to selecting the proper algorithm. Therefore, on the left at Galaxy, we must type the tool's name in the search box. We can change or leave the parameters at the central panel as indicated through the workflow upon selection. Since the *E. coli* libraries were generated using the Illumina platform, it results in short reads (50–300 bases pairs-bp); we need to employ the trimmomatic algorithm (Bolger et al. 2014). For the case of *D. longicaudata*, the libraries were sequenced using the 454 GS FLX Titanium platform, which produced long reads (700 bp). PRINSEQ was implemented in this case (Cantu et al. 2019). It is crucial to investigate the trimming algorithm's characteristics, commonly found in the literature attached to each tool. Besides, splitting the reads into separate files is helpful in cases where the libraries are paired-end reads and not single. This split process can be done with the FASTQ deinterlacer tool (Blankenberg et al. 2010). Once we separate the files, we trim them, and then it is possible to determine the quality of the reads using the FASTQC tool. Thus, we should have essential quality features before moving forward, such as the read sequence length distribution, GC content, and per base sequence quality.

Determine transcript weights at Galaxy

It is possible to utilize different pipelines on the Galaxy platform to analyze our NGS data. Each time we use a tool, the learning machine algorithm automatically suggests logical stepwise tools to guide us in analyzing our data. In these examples, we aim to construct networks to identify genes related to specific conditions like tryptophan metabolism and sex development. Therefore, we decided to treat our data as organisms without a genome of reference and establish the shortest methodology path to determine the expression value of the identified transcripts. Then the route assigned is > de novo contig generation > annotation/blast > weighting the transcript elements.

A contig is a nucleotide sequence assembled from the sequenced nucleic acids. Therefore, its size is larger than the reads that originated it. Because of this, contigs usually harbor different genetic characteristics, such as promoter regions and more than one coding region. Its assembly is performed using "De Bruijn" graphs algorithms, another type of graph theory implementation (Li et al. 2012). Thus, careful attention should be taken when selecting the algorithm to construct contigs de-novo. For the *E. coli* libraries, we should type in the search box on the left of the Galaxy platform RnaSPAdes and select the tool (Bushmanova et al. 2019). Then, we choose the trimmed libraries and execute the algorithm with the default parameters in the middle panel. The QUAST tool determined the quality of the assembled transcripts (Gurevich et al. 2013). Subsequently, it is essential to determine each generated transcript's genetic identity in an annotation process. This step is performed using the PROKKA tool (Seemann 2014). At this point, it is possible, but not necessary, to provide a genome of reference to guide the annotation. The bacteria *E. coli* K 12 genome could be uploaded in FASTA format from NCBI (NC_000913.3) and supplied at the PROKKA tool with the option "Optional FASTA file of trusted proteins"; all the other parameters stay unchanged. For the case of *D. longicaudata* libraries, the process should be the same, but supplying a contig library generated by Mannino et al. (2016). Consequently, we upload the contigs and determine their identity using the MEGABLAST tool (Morgulis et al. 2008) (we recommend searching at the most recent NCBI-NT and RefSeq Genomics). At this point, we have already generated the transcripts de novo or uploaded them and determined their identity.

The final step quantifies the number of reads that construct each transcript at each experimental condition. This process will provide the weight of each unit or gene we assembled and identified at the annotation/blast step. To proceed, we must search for the SAILFISH tool at the left of the Galaxy platform (Rao et al. 2012). This tool is chosen because it requires transcripts, identity, and reads but does not require a genome of reference to determine the number of reads that align with a gene. The algorithm first indexes the transcriptome database by splitting the transcripts into small nucleotide pieces. The tables store the relationships between transcripts and the fractions of nucleotides that construct them. After, the reads are indexed and matched to the transcript tables determining the effective length of the transcript and quantifying the reads associated with them without genome of reference alignment requirements. This metric is expressed in transcripts per million (TPM) (Patro et al. 2014) and would be used to construct the network. Therefore, we need to determine the "transcript weight" of each experimental condition.

For *E. coli*, we will run twice the SAILFISH tool (for minimal media and minimal media + tryptophan) by selecting the genetic identity of the assembled transcript determined by the PROKKA tool with the format FNN at the option "Select the reference transcriptome." Furthermore, in "File containing a mapping of transcripts to genes," the clean reads of the libraries for *E. coli* WT minimal media or *E. coli* WT minimal media + tryptophan. In the case of *D. longicaudata*, we must select the uploaded contigs constructed by Mannino et al. (2016) in the "Select the reference transcriptome" option. Once uploaded the reads (individually each sample), choose the identity of the transcripts determined by MEGABLAST at "File containing a mapping of transcripts to genes." With these approaches, we obtain a database for each experimental condition with the name of the identified transcript, its length, the effective length, TPM, and the number of reads. As stated above, TPM quantifies the relative abundance of each transcript in our samples, and we will consider those data to construct the network. Those data are downloaded at the right (Galaxy platform) panel by clicking on the disk icon on the generated archive. The archive is named "quantification" and not "gene quantification." This archive is in tabular format and can be opened in any spreadsheet by changing the file type extension to CSV. Whit those, it is possible to generate transcription-weighted graphs as described below (Fig. 7).

Construct the network at Cytoscape

Cytoscape is an open-source software platform for visualizing complex networks and integrating these with any type of attribute data (Shannon et al. 2003). Also, the platform offers different applications to perform advanced analysis and modeling. The environment of this platform is divided into three sections (Fig. 6C); on the left, the control panel to fast access to the network view, style editing, and app configuration. The central panel shows a network view. At the bottom is a window with the node/edge/network table with graphical options to extract node or edge information (Fig. 6C). Eventually, a result panel will appear at the right throughout the analyses. Familiarizing with this and other platform options is crucial to reaching optimal performance. We installed two apps to conduct this workflow, (1) the network inference tool to perform network engineering and (2) the clustering tool to identify the structure and dynamics of connected genes in the network ("Analyzing the network" section). These examples were executed in a macOS with 8 GB 1600 MHz DDR3 and 1.6 GHz Dual-Core Intel Core i5.

In a network deconvolution, nodes receive values from transcriptomic data. Edges are inferred based on mathematical algorithms such as statistical or probabilistic methods, Bayesian, Boolean, machine learning, and other ways to determine gene relationships. The Algorithm for the Reconstruction of Accurate Cellular Networks (hereafter ARACNE) identifies candidate interactions by estimating the pairwise gene expression profile of mutual information, coupled with a second step for refining interactions based on data processing inequality (DPI) values (Margolin et al. 2006). This last approach is called network reverse engineering and has been successfully implemented in biological phenomena. We will employ the ARACNE app implemented at the CYNI toolbox plugin at Cytoscape. An external GUI is also available (Margolin et al. 2006).

We need to integrate the TPM counts (see above) assigned to each gene from each experimental condition into one table (Fig. 7). Therefore, we manually merge the spreadsheets (downloaded in "Determine transcript weights at Galaxy" section). For *E. coli*, the TPM values gathered from minimal media and minimal media + tryptophan will be joined to the same spreadsheet but into independent columns at their respective name (transcript-ID). For *D. longicaudata*, we constructed three spreadsheets; larvae-male, larvae-female, and female-male, aiming to identify the transcriptional network dynamics during development. Additionally, we can add an extra column with the name of the genes corresponding to each ID. We obtained the IDs from the MEGABLAST database for *D. longicaudata* or the PROKKA annotation for *E. coli* with the file with extension TSV. Once generated, we will export them into CSV format and produce graphs, as shown in Fig. 7.

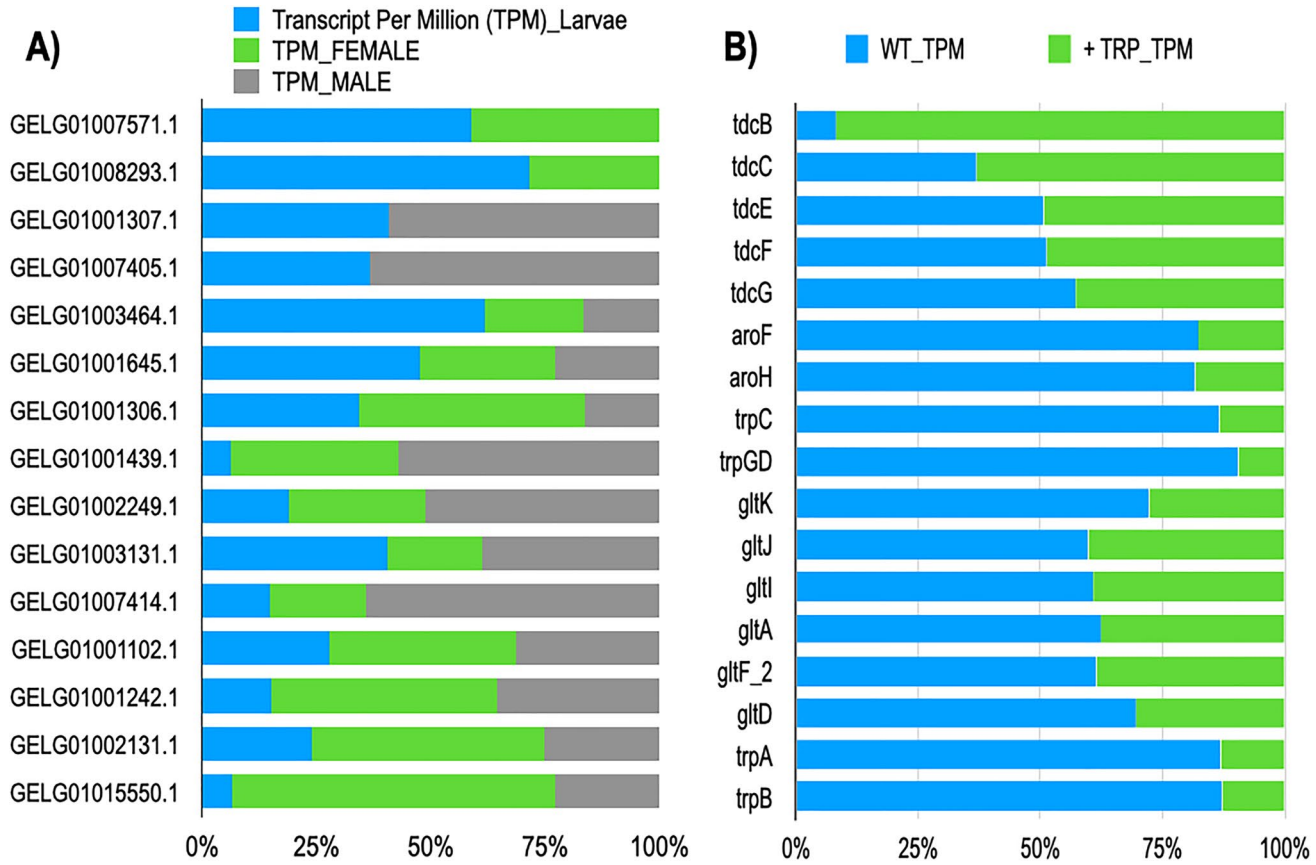

Fig. 7 Transcription-weighted graphs. The spreadsheets harboring the transcript per million (TPM) quantification of reads and mapping to the constructed contigs serve to gather the RNA expression levels between experimental conditions. This database can be sorted or analyzed using any classical approach, such as generating gene expression graphs for **A** *D. longicaudata* or **B** *E. coli*. The different colors represent the TPM value for the indicated gene in the sample. For instance, in **B**, the *trpC* gene is less expressed in *E. coli* when the media is supplemented with tryptophan (green-color bars)

At Cytoscape, we will search for the option to import a network from a file (Fig. 6C). A window will pop up to search and select the database. Upon selection, a new window will show the table with the attributes at the top. In the first column, the name/transcript-ID must be selected to "Source Node-String," and we assigned the TPM values to "Source Node Attribute," other attributes as gene names (or any other added in "Determine transcript weights at Galaxy" section) are unnecessary in this point. A dialogue window will appear upon the "OK" click. Then, click Yes, allowing us to see the nodes as boxes (genes). After saving the project, we can add the remaining attributes from our transcript-weight database that we dismissed initially. We use the "import table from file" option and select the same database. Then, we should choose the name/transcript-ID as "Key-String," and desired attributes are added. With this, we can visualize any feature at the nodes, such as the name of genes. All this information is visualized at the table node at the bottom of the Cytoscape interface.

To construct the network, we should have already installed the CYNI toolbox (Fronczuk et al. 2015) and ARACNE software (as complements apps to Cytoscape). If not, we can do it now at the Apps option. Upon installation, on the left of Cytoscape, we click on the CYNI toolbox and select the proper inference algorithm. Here, we choose the algorithm ARACNE at Discovery mode and Naive Bayes for mutual information determined using the table data we uploaded previously. For instance, the *E. coli* TPM values. Then at the dialogue window of ARACNE, we scroll down, and we make sure that just TPM values are selected at "Source for Network Inference." The algorithm generates networks with 4000 to 7000 nodes and 180,000 to 770,000 edges for *E. coli* and *D. longicaudata*, respectively, by clicking the Apply button. We recommend downloading the yFiles Layout Algorithms App (Wiese et al. 2004) in Cytoscape. This app will allow us to visualize our network on different layouts using minimal computer memory. Therefore, once the network is deconvoluted, we will select the yFiles Organic Layout option at Layout options

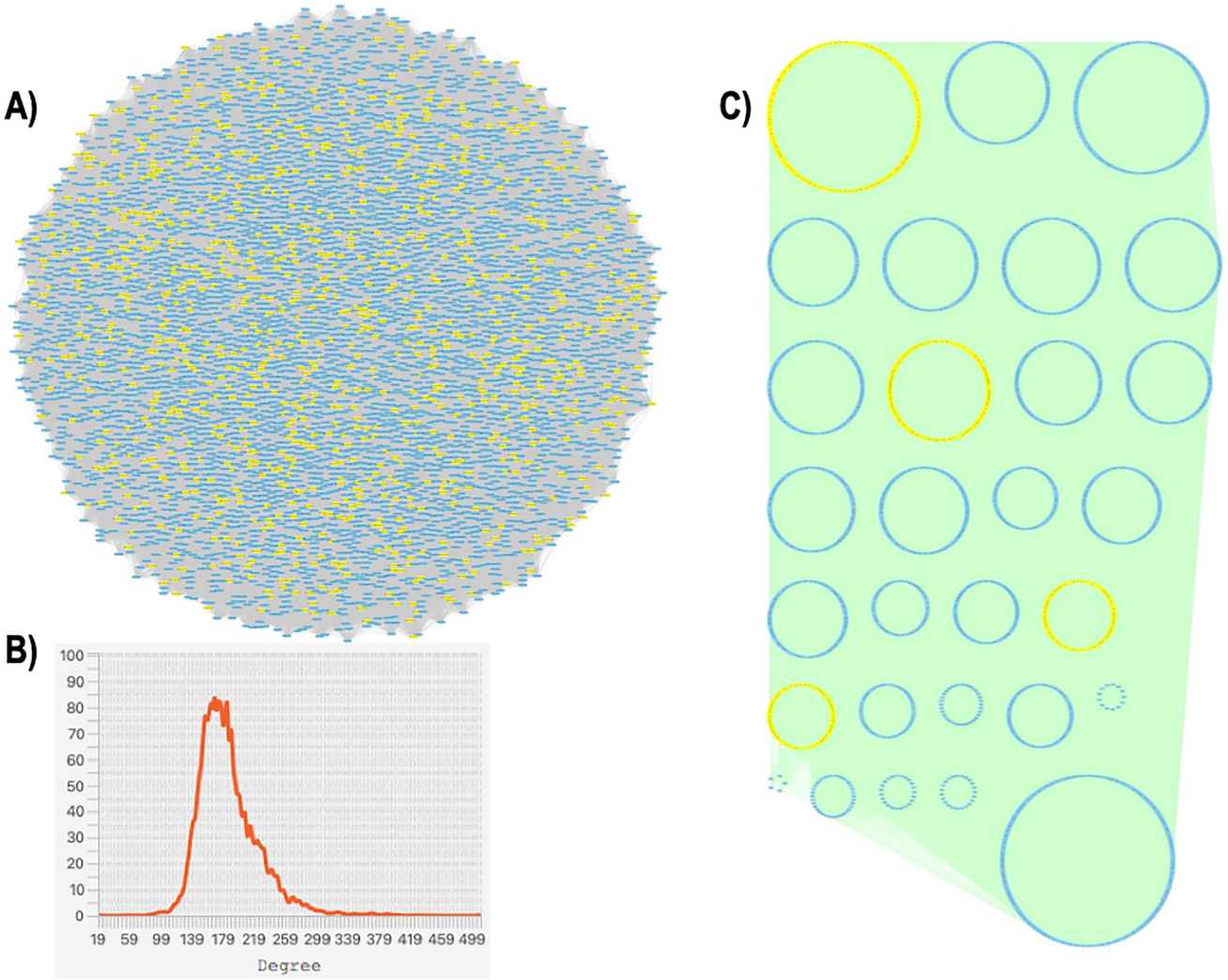

Fig. 8 Network engineering and clustering. We gathered the TPM value from Usergalaxy.UE and manually merged it into a spreadsheet. Then we loaded these data into Cytoscape, and Network inference was performed using ARACNE and clustering using the MCODE algorithm. **A** *D. longicaudata* genetic reconstructed network, **B** *D. longicaudata* node degree distribution histogram, and **C** *E. coli* gene-related tryptophan clusters

to visualize our network. We should get a grid like those pictured in Fig. 8A.

Analyzing the network

The organic layout view suggests a randomized model for our networks. We should provide additional information for the deconvolution process. Then at the tool options, we select "Analyze Network," set the parameters as an undirected graph, and click on "OK." Then, the software will pop up in the analyzer window, and we should click on the node degree distribution on the right. As shown in Fig. 8B, we will obtain one histogram indicative of a randomized network since a curve bell data distribution is observed. Other measurements, such as betweenness and topological coefficient, are added to the table at the bottom of the Cytoscape platform.

Identifying patterns in biological systems is relevant to reveal grouped complexes that putatively perform a task. The clustering algorithms, like ClusterMaker, find densely connected regions in a network using linear algebra (Morris et al. 2011). This clustering offers the advantage of providing the capability of constructing clusters based on connections. Therefore, we must install the ClusterMaker (Morris et al. 2011) and Cytocluster plugins at Cytoscape (Li et al. 2017). These plugins offer a bulk of clustering algorithms. We will work with each app's MCODE (Saito et al. 2012) and ClusterOne (Nepusz et al. 2012) algorithms.

In systems biology, it is demonstrated that a core of about 122 metabolic reactions composed of ~ 130–250 genes are

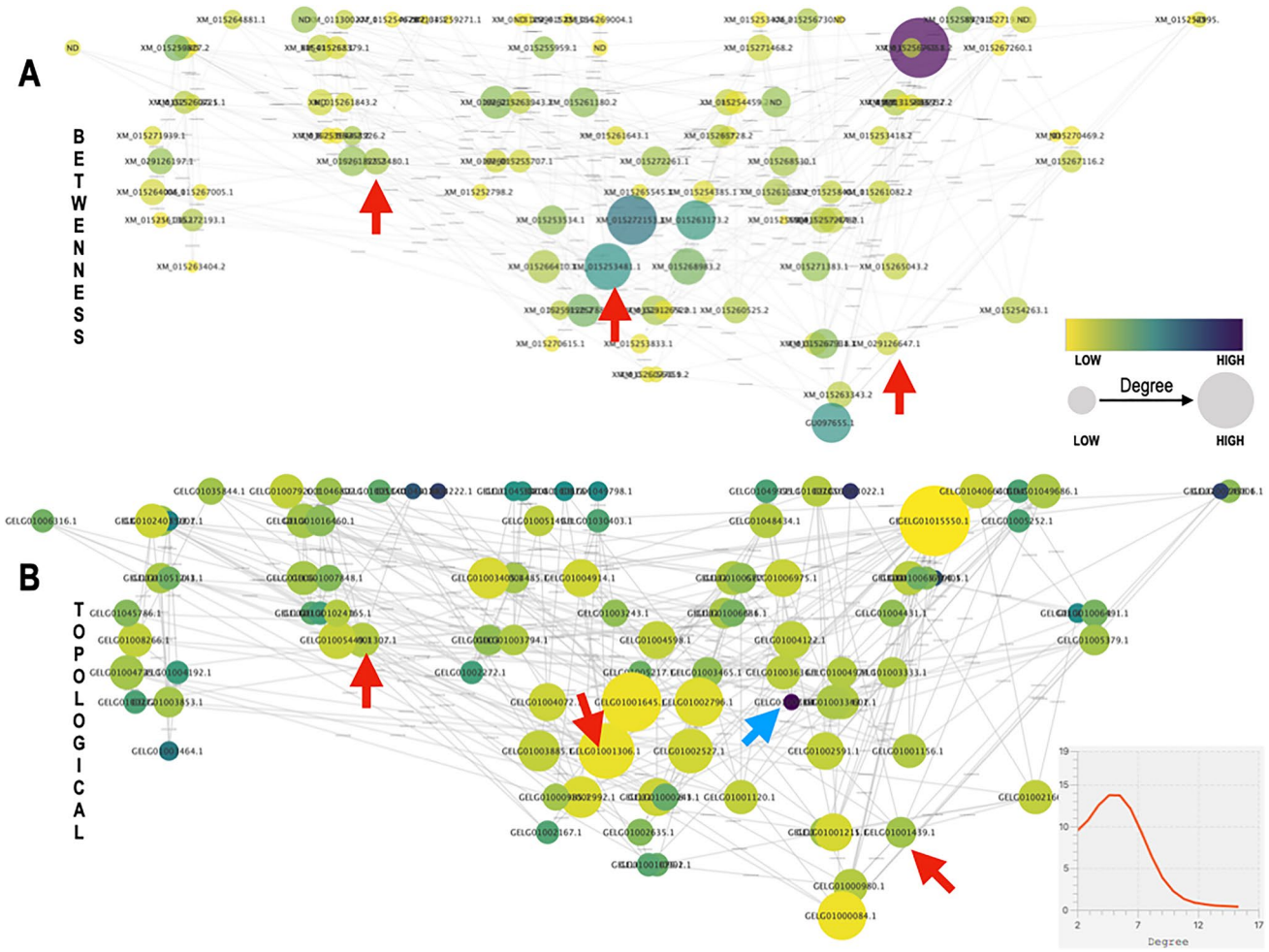

Fig. 9 Centrality measurements in a core sex-related cluster genes of *D. longicaudata* using MCODE. The size of the nodes and their gradient of color represents the degree value in **A** with betweenness and **B** topological analysis results. The red arrows show the known sex-related genes. In (**A**), the betweenness analysis suggests an essential role for the middle node annotated as XM_015253481.1, known in the literature as Transformer-2. The topological network analysis accommodates the nodes with a different relevance

responsible for a biological task (i.e., *E. coli* growth in optimal conditions) (Burgard et al. 2001; Pál et al. 2006). Because of this, we idealize a process for looking at clusters with that average of genes. However, we continue with the process of clustering until we get smaller groups from reduced networks, just aiming to show the capability of this method. Therefore, we execute a reiterative process to find a core of interest-related genes. This process involves the following working flux starting from a given network > clustering process > nodes extraction > engineering network to get "core related-gene of interest." We have already constructed the network using ARACNE in "Construct the network at Cytoscape" section. Now, we will cluster the process as follows: at the Apps option, we select clusterMaker and go to MCODE; we choose this algorithm because it uses the assigned connections which mutual information previously generated from ARACNE inference to determine the clusters, but any other algorithms can also be employed, remember to investigate its parameters before. Upon selection, an MCODE window will appear. We choose the fluff option as it has been proven to predict closer to real connections (Bader and Hogue 2003). All the other parameters must remain unchanged. As soon as the algorithm runs, we will go to the Layout options and select the Group Attributes Layout, "All Nodes," and MCODE Clusters. With this process, we must have results like Fig. 8C for the *E. coli* network. We can search nodes of interest and take the number of the cluster they belong to or search for this at the bottom panel. We also can select the nodes at the active view at the starter panel by clicking on the groups of interest while the ctrl key is pressed (Fig. 8C). Upon selecting all the clusters in which our genes of interest are included, we will export them at the bottom panel of Cytoscape by clicking on the "Export Table to File" option. Then manually exclude

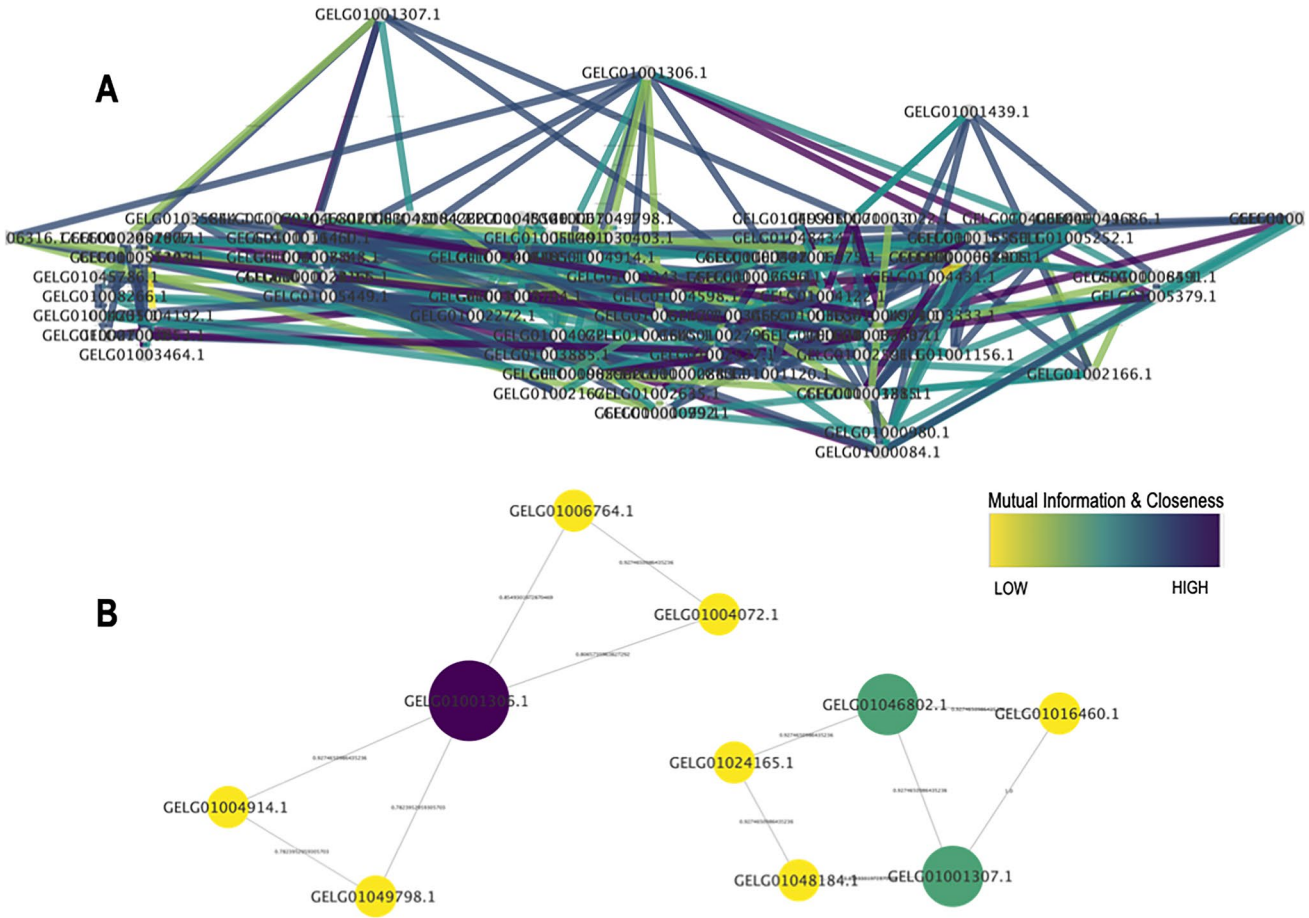

Fig. 10 Clustering approach at the core sex-related cluster genes of *D. longicaudata*. The graph shows the original TPM values from the nodes of the last clustering process using MCODE (reduction from 5955 nodes to 99 using five iterative network reconstructions). The mutual information values were recalculated using ARACNE. **A** We manually moved the sex-related nodes at the top of the hierarchical layout to display the different strengths of the inferred connections and the attributes provided at edge-style options. **B** The clusters related to the genes of interest using the ClusterOne algorithm are also shown

all the generated values in the previous analysis and leave the original TPM values in a spreadsheet. This procedure will import the selected nodes at Cytoscape and repeat the ARACNE inference and the above clustering process. We can export the selected nodes at the current Cytoscape session by clicking "New network from Selection" and then the option "From Selected Nodes, All Edges" (Fig. 6C). The *E. coli* genes to search are those related to the tryptophan operon. For *D. longicaudata*, Megablast provides sex genes such as transformer 2. Depending on our biological question, they should exist genes in which we would be interested. Those are our core genes to select the clusters, which will drive us to the following analyses. With the iterative process of engineering networks and clustering rounds, we can get groups of ~ 150 genes containing those of interest. Those clusters might be optimal to analyze their connections to get possible targets.

Next, we demonstrate the power of the clustering process and network reconstruction to get the minimal network-related clusters related to the inquired biological task. Therefore, for *D. longicaudata*, we generated three networks; those gathering TPM values from larvae-male, larvae-female, and female-male (see above). The weight data were obtained at TPM determination for 51,622 contigs from Mannino et al. (2016), as explained in "Determine transcript weights at Galaxy" section (Fig. 7A). Optionally, and upon computer capabilities, we can decrease the size of the databases by depleting the nodes with values from 0 to 5. The "Construct the network at Cytoscape" section allows generating three networks with 5955 nodes and 529,490 edges for males (Fig. 8A, B). And 7272 nodes with 769,550 edges for the female network and 6125 nodes with 571,482 edges for the adult network. Later we can reduce the number of nodes interacting with sex development genes to 99 nodes with 292 edges

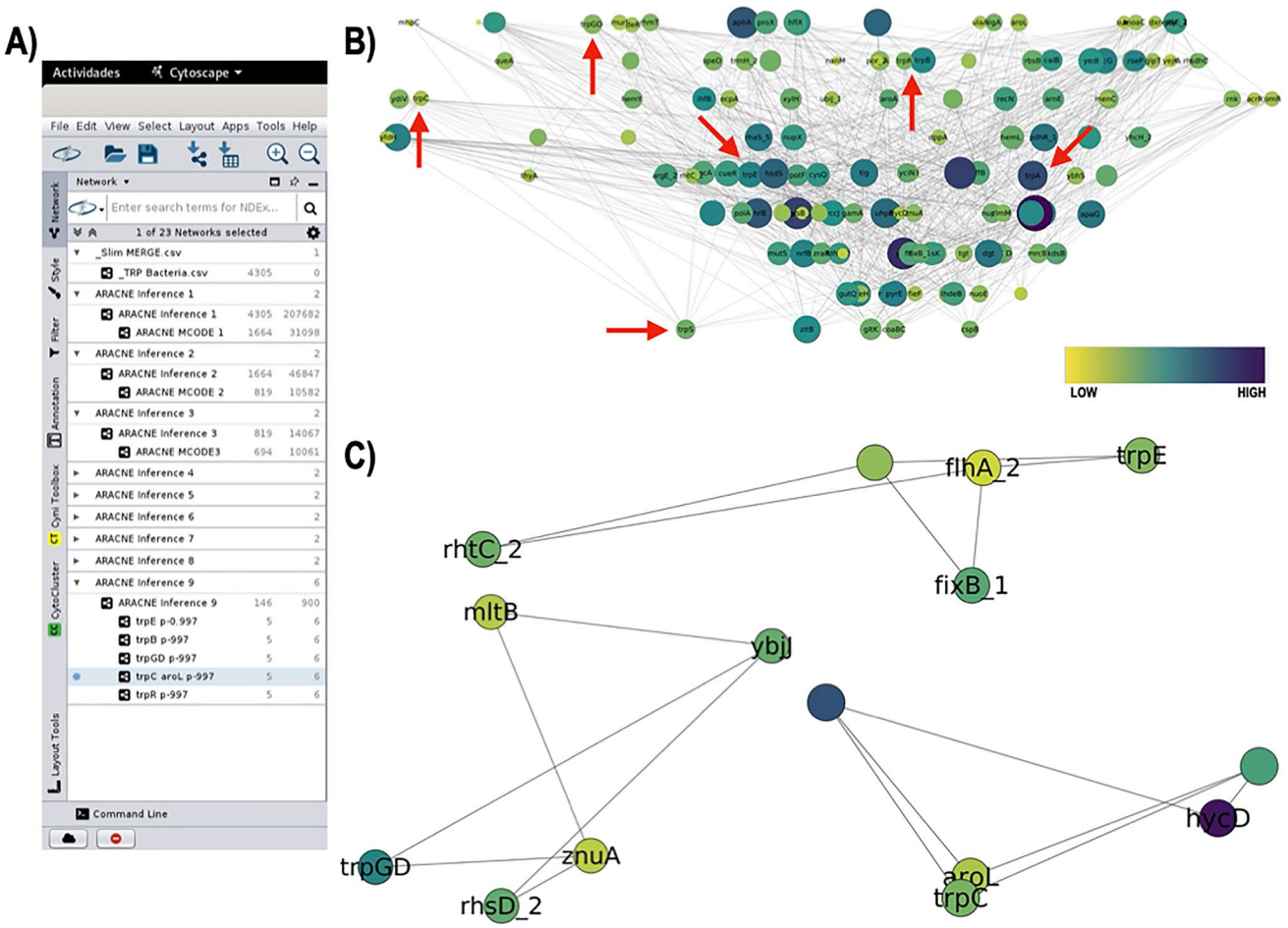

Fig. 11 Tryptophan network analysis on the *E. coli* transcriptome. Results of the complete process described in this work are shown. **A** The control panel screenshot shows the flux of the process, viewing the decrease from 4305 to 146 nodes using network engineering and clustering process. **B** Hierarchical layout (N size) and betweenness analysis (scale-color bar) on the tryptophan core-related genes; red arrows show the trp operon genes. **C** Three clusters obtained from the ClusterOne analysis from the 146 nodes are shown. Some nodes appear without names because they are unannotated

for males, 192 nodes with 760 edges for females, and 74 nodes with 193 edges for adult networks. In Fig. 9, we show the result of this clustering process with subsequent undirected network analysis for the larvae-male network. We display the stylized nodes and edges representing the hierarchical network organization. All the measures are implemented at the control panel in the Style option. The betweenness analysis shows that the sex-related genes, mainly the transformer-2 gene, are among the most important in the network (Fig. 9A). The topological coefficient is a relative measure of how a node shares neighbors with other nodes. This coefficient is something like which of them share friends in common.

The core network sex-related has a high connection, but its topological coefficient is low (Fig. 9B). Outstanding, the extracted data shaped a power-law distribution (Fig. 9 inserted plot), suggesting that the network adopts a scale-free topology, a typical structure for biological networks. We got genes related to mating and neurons conveyed to sex recognition in those clusters. Finally, as an example of the iterative process to get minimal gene collections, we recalculate the mutual information values (Fig. 10A) as a baseline for a new clustering round using the ClusterOne algorithm. We obtain two "core" clusters (Fig. 10B) corresponding to transformer-2 variants X2 and X3, composed of five nodes. The top nodes are related to nucleosome-remodeling factor, E3 ubiquitin-protein ligase, cubilin-like protein, tyrosine-protein kinase CSK, retinal-specific ATP-binding cassette transporter, and a calcium-transporting ATPase. It is important to emphasize that the clustering algorithms working on transcriptomic data consider that there should be more network connections than the knowns. Hence, a deep knowledge of the biological phenomena is essential to guide the network reconstruction, and experimental validation is finally required.

Finally, we repeat to confirm this process using an Ubuntu 20 with 16 GB and 2 GHz Intel Xeon for the *E. coli* data. Figure 11A shows the loop process at Cytoscape of ARACNE inference in complete mode and naive Bayes for MI with MCODE clustering and a final clustering step using ClusterOne. We got a tryptophan core composed of 146 genes. The hierarchical and betweenness analysis suggests that *trpS* and *trpA* are the distinctive nodes in the network (Fig. 11B). In the end, we got five clusters in which *trp* genes are related to other genes. For instance, we retrieve that aromatic amino acid production genes are vital for tryptophan production, such as *aroL*. However, other genes may be unexpected, such as *flhA* (related to motility) and fixB (to carnitine metabolism). However, both are linked to tryptophan metabolism or derivates (Bernal et al. 2007; Li et al. 2019). It is crucial to emphasize that many genes described with a functional role in the tryptophan processing are identified along with the different clustering iterations. Therefore, we recommend a deep analysis of each clustering round before moving forward on functional assays. With these examples, we provide evidence of the capabilities of this pipeline using network theory to identify clusters of genes of interest starting from transcriptomic data.

Conclusion

With the wide availability of genetic information, we encourage the implementation of network theory in studying biological systems. Here, we provide some basic concepts and models employed in network biology analysis, and extensive information can be consulted in the literature. Besides, as examples, we offer a short pipeline based on network theory to process, construct, and analyze transcriptome data for two biological processes. With the big data available, it is essential to have prospective tools to generate working hypotheses that can be validated experimentally. Better exploitation of these platforms and analysis tools is achieved by familiarizing them and getting some computation capabilities. The careful research and well-designed hypothesis are critical hallmarks to maximally exploiting the network theory capabilities.

Acknowledgements The authors thank the Laboratory for Learning and Research in Biological Computing at the Center for Research and Advanced Studies of the National Polytechnic Institute (Laicbio, Cinvestav Irapuato) for providing a computer with Ubuntu 20 with 16 GB and 2 GHz Intel Xeon to conduct part of this research.

Author contribution A. M-A and G. R-A conceived the project, analyzed the data, and wrote and edited the manuscript.

Funding This research was funded by CONACYT 2022 (grant 319732 given to AM-A).

Data availability The datasets analyzed in this work were accessed at the GenBank SRR database https://ncbi.nlm.nih.gov/ and are available with the accession numbers SRR922261 and SRR922264 for *E. coli* library and Bio project PRJNA317427-GELG01.1, SRR3336273.1, SRR3336336.1, and SRR3336337.1 for *D. longicaudata* database.

Declarations

Competing interest The authors declare no conflict of interest.

References

Alm E, Arkin AP (2003) Biological networks. Curr Opin Struct Biol 13(2):193–202. https://doi.org/10.1016/s0959-440x(03)00031-9

Amaral LAN, Ottino JM (2004) Complex networks. Eur Phys J B 2(38):147–162. https://doi.org/10.1140/epjb/e2004-00110-5

Amores GR, de Las Heras A, Sanches-Medeiros A, Elfick A, Silva-Rocha R (2017) Systematic identification of novel regulatory interactions controlling biofilm formation in the bacterium Escherichia coli. Sci Rep 7(1):16768. https://doi.org/10.1038/s41598-017-17114-6

Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4:2. https://doi.org/10.1186/1471-2105-4-2

Barabási AL (2002) The new science of networks. Cambridge MA. Perseus. Am J Phys 71:409. https://doi.org/10.1119/1.1538577

Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512. https://doi.org/10.1126/science.286.5439.509

Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2):101–113. https://doi.org/10.1038/nrg1272

Barnes JA, Harary F (1983) Graph theory in network analysis. Social Netw 5(2):235–244. https://doi.org/10.1016/0378-8733(83)90026-6

Bernal V, Sevilla A, Cánovas M, Iborra JL (2007) Production of L-carnitine by secondary metabolism of bacteria. Microb Cell Fact 6:31. https://doi.org/10.1186/1475-2859-6-31

Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A, Galaxy Team (2010) Manipulation of FASTQ data with Galaxy. Bioinformatics 26(14):1783–1785. https://doi.org/10.1093/bioinformatics/btq281

Bohra A, Rathore A, Gandham P, Saxena RK, Satheesh Naik SJ, Dutta D, ... , Singh N P (2021) Genome-wide comparative transcriptome analysis of the A4-CMS line ICPA 2043 and its maintainer ICPB 2043 during the floral bud development of pigeonpea. Funct Integr Genomics, 21(2), 251-263.https://doi.org/10.1007/s10142-021-00775-y

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bordbar A, Nagarajan H, Lewis NE, Latif H, Ebrahim A, Federowicz S, Schellenberger J, Palsson BO (2014) Minimal metabolic pathway structure is consistent with associated biomolecular interactions. Mol Syst Biol 10(7):737. https://doi.org/10.15252/msb.20145243

Burgard AP, Vaidyaraman S, Maranas CD (2001) Minimal reaction sets for Escherichia coli metabolism under different growth requirements and uptake environments. Biotechnol Prog 17(5):791–797. https://doi.org/10.1021/bp0100880

Bushmanova E, Antipov D, Lapidus A, Prjibelski AD (2019) rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience 8(9):giz100. https://doi.org/10.1093/gigascience/giz100

Cantu VA, Sadural J, Edwards R (2019) PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets. PeerJ Preprints 7:e27553v1. https://doi.org/10.7287/peerj.preprints.27553v1

Carabajal Paladino L, Muntaabski I, Lanzavecchia S, Le Bagousse-Pinguet Y, Viscarret M, Juri M, Fueyo-Sánchez L, Papeschi A, Cladera J, Bressa MJ (2015) Complementary sex determination in the parasitic wasp Diachasmimorpha longicaudata. PLoS ONE 10(3):e0119619. https://doi.org/10.1371/journal.pone.0119619

Cheah PL, Looi LM (2001) p53: an overview of over two decades of study. Malays J Pathol 23(1):9–16

Crawford IP (1975) Gene rearrangements in the evolution of the tryptophan synthetic pathway. Bacteriol Rev 39(2):87–120. https://doi.org/10.1128/br.39.2.87-120.1975

de Sola Pool I, Kochen M (1978) Contacts and influence. Social Netw 1(1):5–51. https://doi.org/10.1016/0378-8733(78)90011-4

Erdos P, Rényi A (1960) On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5(1):17–60. https://snap.stanford.edu/class/cs224w-readings/erdos60random.pdf. Accessed 1 Oct 2022

Estrada E (2013) Graph and network theory. Mathematical Tools for Physicists. 2nd Edition (editor: M. Grinfeld). John Wiley & Sons. https://doi.org/10.1002/3527600434.eap726

Fronczuk M, Raftery AE, Yeung KY (2015) CyNetworkBMA: a Cytoscape app for inferring gene regulatory networks. Source Code Biol Med 10(1):1–7. https://doi.org/10.1186/s13029-015-0043-5

Goecks J, Nekrutenko A, Taylor J, Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology 11(8):R86. https://doi.org/10.1186/gb-2010-11-8-r86

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29(8):1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Itzkovitz S, Alon U (2005 Subgraphs and network motifs in geometric networks. Physical review. E, Statistical, nonlinear, and soft matter physics, 71(2 Pt 2), 026117. https://doi.org/10.1103/PhysRevE.71.026117

Li M, Li D, Tang Y, Wu F, Wang J (2017) CytoCluster: a cytoscape plugin for cluster analysis and visualization of biological networks. Int J Mol Sci 18(9):1880. https://doi.org/10.3390/ijms18091880

Li Y, Liu B, Guo J, Cong H, He S, Zhou H, Zhu F, Wang Q, Zhang L (2019) L-Tryptophan represses persister formation via inhibiting bacterial motility and promoting antibiotic absorption. Future Microbiol 14:757–771. https://doi.org/10.2217/fmb-2019-0051

Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, ... Fan W (2012) Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. Brief Funct Genom 11(1):25-37.https://doi.org/10.1093/bfgp/elr035

Mannino MC, Rivarola M, Scannapieco AC, González S, Farber M, Cladera JL, Lanzavecchia SB (2016) Transcriptome profiling of Diachasmimorpha longicaudata towards useful molecular tools for population management. BMC Genom 17(1):793. https://doi.org/10.1186/s12864-016-2759-2

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinform 7(Suppl 1):S7. https://doi.org/10.1186/1471-2105-7-S1-S7.10.1186/1471-2105-7-S1-S7

Martínez-Antonio A, Collado-Vides J (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. Curr Opin Microbiol 6(5):482–489. https://doi.org/10.1016/j.mib.2003.09.002

Milgram S (1967) The small world problem. Psychol Today, 2(1):60–67. https://courses.cit.cornell.edu/info2950_2012sp/milgram.pdf

Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA (2008) Database indexing for production MegaBLAST searches. Bioinformatics 24(16):1757–1764. https://doi.org/10.1093/bioinformatics/btn554

Morris JH, Apeltsin L, Newman AM, Baumbach J, Wittkop T, Su G, Bader GD, Ferrin TE (2011) clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinform 12:436. https://doi.org/10.1186/1471-2105-12-436

Nepusz T, Yu H, Paccanaro A (2012) Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods 9(5):471–472. https://doi.org/10.1038/nmeth.1938

Pál C, Papp B, Lercher MJ, Csermely P, Oliver SG, Hurst LD (2006) Chance and necessity in the evolution of minimal metabolic networks. Nature 440(7084):667–670. https://doi.org/10.1038/nature04568

Palsson B (2015) Systems biology: constraint-based reconstruction and analysis. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139854610

Patro R, Mount SM, Kingsford C (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol 32(5):462–464. https://doi.org/10.1038/nbt.2862

Rao S, Ramakrishnan R, Silberstein A, Ovsiannikov M, Reeves D (2012) Sailfish: a framework for large scale data processing. In Proceedings of the Third ACM Symposium on Cloud Computing. pp. 1–14. https://doi.org/10.1145/2391229.2391233

Ravasz E, Barabási AL (2003) Hierarchical organization in complex networks. Physical review. E, Statistical, nonlinear, and soft matter physics, 67(2 Pt 2):026112. https://doi.org/10.1103/PhysRevE.67.026112

Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, ... , Ideker T (2012) A travel guide to Cytoscape plugins. Nat Methods 9(11):1069-1076.https://doi.org/10.1038/nmeth.2212

Sayers EW, Beck J, Bolton EE, Bourexis D, Brister JR, Canese K, Comeau DC, Funk K, Kim S, Klimke W, Marchler-Bauer A, Landrum M, Lathrop S, Lu Z, Madden TL, O'Leary N, Phan L, Rangwala SH, Schneider VA, Skripchenko Y, … Sherry ST (2021) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 49(D1):D10–D17. https://doi.org/10.1093/nar/gkaa892

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30(14):2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504. https://doi.org/10.1101/gr.1239303

Taylor J, Schenck I, Blankenberg D, Nekrutenko A (2007) Using Galaxy to perform large-scale interactive data analyses. Current protocols in bioinformatics, Chapter 10, Unit–10.5. https://doi.org/10.18632/oncotarget.20488

Wiese R, Eiglsperger M, Kaufmann M (2004) yfiles—visualization and automatic layout of graphs. In Graph Drawing Software (pp. 173–191). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-18638-7_8

Yu D, Kim M, Xiao G, Hwang TH (2013) Review of biological network data and its applications. Genom Inform 11(4):200–210. https://doi.org/10.5808/GI.2013.11.4.200

Zaytsev E (2008) Euler's Problem of Königsberg Bridges and Leibniz'Geometria Situs. Arch Int Hist Sci 58(160–161):151–170. https://doi.org/10.1484/J.ARIHS.5.101505