SAMBA: Structure-Learning of Aquaculture Microbiomes Using a Bayesian-Network Approach

Beatriz Soriano[1,2,3*], Ahmed Hafez[1,4], Fernando Naya-Català[2], Roxana Andreea Moldovan[1], Socorro Toxqui-Rodríguez[2], M. Carla Piazzon[2], Vicente Arnau[3], Carlos Llorens[1], Jaume Perez-Sánchez[2*]

[1]Biotechvana, Parc Científic Universitat de València, Spain

[2]Institute of Aquaculture Torre de la Sal (IATS, CSIC), Castellón, Spain

[3]Institute for Integrative Systems Biology (I2SysBio), UVEG-CSIC, Valencia, Spain.

[4]Faculty of Computers and Information, Minia University, Egypt

*Corresponding authors: beatriz.soriano@biotechvana.com; jaime.perez.sanchez@csic.es

Abstract

Summary: Like in other animals, the gut microbiome of fishes contains thousands of microbial species that establish a complex network of relationships among each other and with the host. These interrelationships are shaped by biotic and abiotic factors, but little is known about how they evolved and how they are regulated by the environment, farmers and breeders. Herein, we introduce SAMBA (Structure-Learning of Aquaculture Microbiomes Using a Bayesian-Network Approach), the software implementation of a Bayesian network model for investigating how fish pan-microbiomes and all other variables of a given aquaculture system are related each to other. SAMBA is powered by a Bayesian network trainable model that learns the network structure of an aquaculture system using information from distinct biotic and abiotic variables of importance in fish farming, with special focus on microbial data provided from 16S amplicon sequencing. SAMBA accepts both qualitative and quantitative variables and convincingly deals with the differences in microbial composition derived by the technical or biological variation among microbiomes of distinct specimens. To this end, SAMBA is implemented with a variety of tools to pre-analyze the data and chose a distribution (log-normal, binomial, etc.) to build and train the Bayesian network model. Once the model has been created and validated, the user can interrogate the model and obtain information about the modelled system in two different modes: Report and Prediction. Using the Report mode SAMBA reports how the pan-microbiome and all other variables involved in the modelled aquaculture system influence each other and what are the conditional probabilities of each relation. Under the Prediction mode, the application predicts how the diversity and functional profile of the pan-microbiome would likely change depending on any change made on other variables. Finally, SAMBA implements a comprehensive graphical network editor allowing to navigate, edit and export outcomes. We tested and validated the performance of SAMBA using microbial community standards and gut microbiota communities of farmed gilthead sea bream (*Sparus aurata*) from different feeding trials.

Availability and implementation: SAMBA is a web-based solution for network analysis. The backend component is a workflow implemented in R and Python using third party software dependencies while the front end component is the environment of graphical user interfaces implemented to manage SAMBA. The source code of SAMBA, the datasets to test it and instructions to execute the application are available at https://github.com/biotechvana/SAMBA.

Contact: beatriz.soriano@biotechvana.com, jaime.perez.sanchez@csic.es

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1. Background

Fish gut bacterial communities are subjected to large fluctuations that reflect the complex dialogue among the host and resident microbial communities. Such interactions are driven by a number of intrinsic (e.g. genotype, physiological status, pathobiology) and environmental factors (e.g. environment, lifestyle and diet) that have the capacity to modify the composition and diversity of gut microbiota as well as its function and metabolic activity (Egerton et al., 2018). Indeed, current research in farmed gilthead sea bream (*Sparus aurata*) highlighted the use of gut microbiota as a reliable criterion to evaluate the success of selective breeding to produce healthier and more resilient fish with changes in diet composition and limited availability of conventional feedstuffs (Piazzon et al., 2020; Naya-Català et al., 2022a; 2022b). However, our understanding about the dynamics of interrelations between fish microbiomes and their hosts is still in its infancy due to the multiple biotic and abiotic factors involved in this kind of dynamics (Faust, 2021; Liu et al., 2021). Certainly, there is now a well-known recognized interest for system biology tools like Bayesian networks (BN) and structure learning systems (Scutari, 2009; Michiels et al., 2021) that can be convincingly used to realistically modelize complex biological systems. In the field of Microbiology, BNs have proven to be a powerful tool for inferring directional relationships within microbial communities (Sazal et al., 2020; Sazal et al., 2021), and for analyzing functional networks in metagenomics data (Hobbs et al., 2016). Current state-of-the-art in BNs includes several tools with interactive visualizations, such as *ShinyBN* (Chen et al., 2019) or *BayesiaLab* (Conrady and Jouffe, 2015), which are only able to work with discrete variables and small datasets. The recent *BayeSuites* tool (Michiels et al., 2021) overcomes these difficulties, being capable of managing continuous variables and large datasets focused on neurosciences, albeit it is not able to perform inference on discrete variables to establish their conditional probability distributions. In aquaculture research, BNs have been used to better integrate knowledge on sustainable ecosystem-based management (Yuniarti et al., 2021), but the structure of interrelations and dependencies among the multiple distinct biotic and abiotic factors

involved in the dynamics of a given aquaculture system remains mostly unknown. The objective of this work is to present SAMBA (Structure-Learning of Aquaculture Microbiomes using a Bayesian-Network Approach), the software implementation of a BN model allowing to investigate and/or predict how fish pan-microbiomes and other variables involved in the dynamics of given aquaculture system are related and influence each other.

2. Results

2.1. General overview

SAMBA is a web-based application that reads and indexes data from distinct biotic and abiotic variables involved in the dynamics of a given aquaculture system to learn the network structure of that system and create and train a BN model based on that information. Typical variables of importance in aquaculture systems to build a BN model with SAMBA are sex, age, size, genetic background, tissue, microbial frequencies (taxonomic annotations from 16S amplicon sequencing), season, diet composition, pH, temperature, etc. Next, SAMBA infers the conditional probabilities of all possible interrelations of dependency detected among these variables indexed in the BN model with special focus on the distinct taxa constituting the fish pan-microbiome of the modelled system, from which SAMBA also permits to infer a metagenome profile. Once the BN model has been created, trained, and validated, SAMBA permits the user to infer and/or extract information of the network model as a report or as a prediction. In addition, SAMBA implements a comprehensive graphical network editor that provides multiple distinct tools to visualize, edit, customize, and export networks in distinct formats.

2.2. Implementation and usage

SAMBA has been implemented as an interactive web application using R, Python and Shiny (Chang et al., 2021). As shown in Figure 1, this GUI is structured in five modules designated "Input Network", "Network Reports", "Prediction", "Network Viewer" and "Results" that provide interface

implementations to the distinct SAMBA modules. A brief description of each module is provided next:

2.2.1. Input Network

This module represents the first step of the SAMBA's workflow. It is composed of three different sections: *"Load Network"*, where a BN model can be uploaded to perform further analysis in SAMBA; "*Evidence/Control*", where SAMBA reports a summary table for each taxon and a graph showing the values distribution in one condition or two different conditions if the user selects the options of comparing them; and "*Learning and training*" which reads the input data from a csv file and build the BN model using the *bnlearn* R package (Scutari, 2010) with the score-based learning Hill-Climbing algorithm (Selman and Gomes, 2006). To deal with the technical and biological variability and sparsity among microbiomes of distinct specimens (Faust, 2021), this module implements the following tools: i) filtering options to remove taxa present in few samples (prevalence filter); ii) white- or black-lists to establish or forbid causal relationships between nodes based on prior knowledge or some assumption about the underlying biological model; and iii) distinct score functions to be used in the Hill-Climbing learning algorithm. Currently, SAMBA provides the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the log-likelihood (loglik) scores. The calculation of these scores can be done using a log-normal distribution provided by *bnlearn* and a zero-inflated negative binomial distribution (ZINB) provided by *pscl* (Zeileis et al., 2008). Once the data has been loaded and filtered, the user can train the BN model to infer the conditional dependencies among all the variables. The significance and strength of the global model and each of the relationships are calculated using the BIC and Mutual Information (MI) criterion (De Campos, 2006).

2.2.2. Network Reports

Once the BN model has been created and trained, this module uses *bnlearn* and *dagitty* (Textor et al. 2016) to obtain key information about the potential causality and relationships among microbial taxa and all other variables involved in the dynamics of the modelled aquaculture species. In particular, the user can obtain a report with metrics and statistics with information of network structure of the modelled system, including conditional probability tables (CPT), conditional probability values (CPV), and a residuals and Q-Q plot for each node (i.e. each indexed variable).

2.2.3. Prediction

This module uses *bnlearn* and *pscl* to infer predictions about the most likely changes on the diversity and frequencies of the pan-microbiome depending on changes in other variables. This module also permits to infer the functional metagenomic profile associable to the modelled pan-microbiome using PICRUSt2 (Douglas et al., 2020). In addition, the module permits the user to change the prediction according to the user's experimental expertise on the modelled system thus allowing to calibrate the network predictions until they converge with the real-world observation.

2.2.4. Network Viewer

This module is the implementation of a comprehensive graphical network editor based on distinct functions of the following R packages: *igraph* (Csardi and Nepusz, 2006), *visNetwork* (Almende et al., 2019), *bnviewer* (Fernandes, 2020), *shinythemes* (Chang et al., 2021), *shinyBS* (Bailey, 2022), *tibble* (Muller and Wickham, 2022), *colourpicker* (Attali, 2022), shinydashboard (Chang et al., 2021), shinydashboardPlus (Granjon, 2021) shinyscreenshot (Attali et al., 2021), shinyjs (Attali, 2021), htmlwidgets (Vaidyanathan et al., 2022), stringr (Wickham, 2022), purrr (Henry and Wickham, 2021), DT (Yihui et al., 2022). This network editor provides multiple distinct tools to visualize, edit, customize, and export networks in distinct modes and formats, facilitating navigation and interpretation of the graph (or a subgraph) resulting from the BN model.

2.2.5. Results

This module is the section where SAMBA stores the results from where the user can download the output files generated from the Learning and training interface, including normalized counts, link strengths and an RData file containing the BN model, and the output files generated from the metagenome inference in the Prediction interface.

The whole implementation integrates additional functions of the *Future* package (Bengtsson, 2021) to execute the learning and training, and prediction steps in the background allowing other tasks of SAMBA to work while the aforesaid steps are still running. For more detailed information about the frontend and Backend implementation and usage of SAMBA, please refer to the Supplementary Data file accompanying this article. Therein, we also provide detailed information about the algorithmic basis of the BN model, structure learning, data normalization and training methods.

2.3. Testing datasets and validation

SAMBA has been successfully validated using distinct empirical datasets. First, we used three replicates of the ZymoBIOMICS^{TM} Microbial Community Standard II (Zymo Research Corp., CA, United States). This synthetic population is composed by bacterial species whose concentrations are defined and logarithmically distributed. Thus, there is only technical variability within replicates, allowing elucidation of the relationships and their associated taxa predicted values (Supplementary Data File, Section Case-Study Validation). Second, we used a pan-microbiome dataset derived from three feeding trials with the same starting point (May 2020) and location (40°5′N; 0°10′E) and including sibling fish purchased at the same Mediterranean hatchery (Piscimar, Burriana, Spain) (Naya-Català et al., 2021; Solé-Jiménez et al., 2021; Piazzon et al., 2022). These three datasets derive from the real-world observation and were thus important to test how SAMBA manages the technical and/or biological variability among microbiomes of distinct specimens. Predictions performed with SAMBA based on these three datasets indicate that the microbial variability do not affect the interrelations detected among the modelled variables or the direction of these interrelations. Such

variability negatively influences the reliability of taxa frequencies that are inferred when the values of other variables change (for example diet composition), but the percentage of success in predicting the most likely range of frequencies for each taxa is higher than 90% (Supplementary Data, Section Case-Study Validation). This is because SAMBA is able to detect how the input data is distributed, and with that information, the application allows the users to choose different options to build the BN model according to the most appropriate distribution. The taxa frequencies are predicted in ranges with distinct assigned probabilities, and SAMBA provides additional tools for filtering the less prevalent taxa or for editing the frequencies and their CPTs according to the user knowledge to re-run the predictions until they converge with the real-world observation.

3. Concluding remarks

Modelling the complex relationships between the physical and biological components of aquaculture systems is one of the future challenges in the field of food production in the context of climate change and human population growth (Abberton et al., 2016; Poore and Nemecek, 2018). Here, we have introduced SAMBA, the software implementation of a BN model with multiple practical applicabilities. SAMBA may help to reveal the relationships and dependencies within a pan-microbiome, identifying the taxa that constitute the core pan-microbiome and determine their order of colonization. This information can be of particular interest in fish farming, because the dynamics and hierarchies of the fish pan-microbiomes of a given aquaculture system can be used to study the effects of dietary formulations, environmental stimulus, or traceability factors, all of them contributing to improve the sustainability of the aquaculture industry (Terova et al., 2022). In general, SAMBA is very intuitive and easy to use granted to its GUI implementation, and it can be considered as a project in continuous progress. We are committed to implement new functions and tools in further releases of SAMBA, for example new methods of normalization, new data distributions and/or new functions to detect the minimum training sample size for a given

experiment, which may help to increase the prediction power of BNs (see for example, Hu et al., 2021). In summary, SAMBA permits to infer potential relationships within the entities constituting a microbial community, helping to understand how these communities establish causal dependencies among them in the mucosal barriers of fish. Finally, it is worth to stress that, although we have tested and validated SAMBA using data from famed fish, the application can be used to model any kind of microbiome-host network relationships using as case study any other complex organism, including humans.

Funding

This work was supported by the Spanish MCIN project Bream-AquaINTECH (RTI2018–094128-B-I00, AEI/FEDER, UE) to JP-S and JC-G. This study also forms part of the ThinkInAzul programme and was supported by MCIN with funding from European Union NextGenerationEU (PRTR-C17.I1) and by Generalitat Valenciana (THINKINAZUL/2021/024) to JP-S. BS was supported by a pre-doctoral research fellowship from Industrial Doctorate of MINECO (DI-17-09134). FN-C was supported by a research contract from the EU H2020 Research Innovation Program under grant agreement no. 818367 (AquaIMPACT). MCP was funded by a Ramón y Cajal Postdoctoral Research Fellowship (RYC2018-024049-I co-funded by AEI, European Social Fund (ESF) and ACOND/2022 Generalitat Valenciana).

References

Abberton, M. et al. (2016) Global agricultural intensification during climate change: a role for genomics. Plant Biotechnol. J., 14, 1095–1098.

Almende, B.V. et al. (2019) visNetwork: Network visualization using 'vis.js' library. R package version 2.0.9. https://CRAN.R-project.org/package=visNetwork.

Attali, D. (2022) colourpicker: A Colour Picker Tool for Shiny and for Selecting Colours in Plots. R package version 1.2.0. https://CRAN.R-project.org/package=colourpicker.

Attali, D. (2021) shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds. R package version 2.1.0. https://cran.r-project.org/web/packages/shinyjs/index.html.

Attali, D. et al. (2021) shinyscreenshot: Capture Screenshots of Entire Pages or Parts of Pages in 'Shiny'. R package version 0.2.0. https://cran.r-project.org/web/packages/shinyscreenshot/index.html.

Bailey, E. (2022) shinyBS: Twitter Bootstrap Components for Shiny. R package version 0.61.1. https://CRAN.R-project.org/package=shinyBS.

Bengtsson, H. (2021) A Unifying Framework for Parallel and Distributes Processing in R using Futures. R J., 13, 273-291.

Chang, W. et al. (2021) shiny: web application framework for r. R package version 1.6.0. https://CRAN.R-project.org/package=shiny.

Chang, W. (2021) shinythemes: shinythemes: Themes for Shiny. R package version 1.2.0. https://CRAN.R-project.org/package=shinythemes.

Chang, W. and Borges, B. (2021) shinydashboard: Create Dashboards with 'Shiny'. R package version 0.7.2. https://cran.r-project.org/web/packages/shinydashboard/index.html.

Chen, J. et al. (2019) shinyBN: an online application for interactive Bayesian network inference and visualization. BMC Bioinform., 20, 711.

Conrady, S. and Jouffe, L. (2015) Bayesian networks and BayesiaLab: a practical introduction for researchers (Vol. 9). Franklin: Bayesia USA.

Csardi, G. and Nepusz, T. (2006) The igraph software package for complex network research. InterJournal, Complex Systems, 1695.

De Campos, L.M. (2006) A scoring function for learning bayesian networks based on mutual information and conditional independence tests. J. Mach. Learn. Res., 7, 2149–2187.

Douglas, G.M. et al. (2020) PICRUSt2 for prediction of metagenome functions. Nat. Biotechnol., 38, 685–688.

Egerton, S. et al. (2018) The Gut Microbiota of Marine Fish. Front. Microbiol., 9, 873.

Faust, K. (2021) Open challenges for microbial network construction and analysis. ISME J., 15, 3111–3118.

Fernandes, R. (2020) Bnviewer: Bayesian networks interactive visualization and explainable artificial intelligence. R package version 0.1.6. https://CRAN.R-project.org/package=bnviewer.

Granjon, D. (2021) shinydashboardPlus: Add More 'AdminLTE2' Components to 'shinydashboard'. R package version 2.0.3. https://cran.r-project.org/web/packages/shinydashboardPlus/index.html.

Henry, L. and Wickham, H. (2021) purrr: Functional Programming Tools. R package version 0.3.5. https://cran.r-project.org/web/packages/purrr/.

Hobbs, E.T. et al. (2016) A Bayesian inference method for the analysis of transcriptional regulatory networks in metagenomic data. Algorithms Mol. Biol., 11, 19.

Hu, J. et al. (2021) Minimum training sample size requirements for achieving high prediction accuracy with the BN model: A case study regarding seismic liquefaction. Expert Syst. Appl., 185, 115702.

Liu, Z. et al. (2021) Network analyses in microbiome based on high-throughput multi-omics data. Brief. Bioinform., 22, 1639–1655.

Michiels, M. et al. (2021) BayeSuites: An open web framework for massive Bayesian networks focused on neuroscience. Neurocomputing, 428, 166–181.

Muller, K. and Wickham, H. (2022) tibble: Simple Data Frames. R package version 3.1.8. https://CRAN.R-project.org/package=tibble.

Naya-Català, F. et al. (2021) Modulation of gilthead sea bream gut microbiota by a bioactive egg white hydrolysate: interactions between bacteria and host lipid metabolism. Front. Mar. Sci., 8, 698484.

Naya-Català, F. et al. (2022a) Diet and host genetics drive the bacterial and fungal intestinal metatranscriptome of gilthead sea bream. Front. Microbiol., 13, 883738.

Naya-Català, F. et al. (2022b). Genetics and Nutrition Drive the Gut Microbiota Succession and Host-Transcriptome Interactions through the Gilthead Sea Bream (*Sparus aurata*) Production Cycle. Biology, 12, 1744.

Piazzon, M.C. et al. (2020) Genetic selection for growth drives differences in intestinal microbiota composition and parasite disease resistance in gilthead sea bream. Microbiome, 8, 168.

Piazzon, M.C. et al. (2022). A novel fish meal-free diet formulation supports proper growth and does not impair intestinal parasite susceptibility in gilthead sea bream (*Sparus aurata*) with a reshape of gut microbiota and tissue-specific gene expression patterns. Aquaculture, 158, 738362.

Poore, J. and Nemecek, T. (2018) Reducing Food's Environmental Impacts Through Producers and Consumers. Science, 360, 987–992.

Sazal, M. et al. (2020) Inferring directional relationships in microbial communities using signed Bayesian networks. BMC Genom., 21, 663.

Sazal, M. et al. (2021) Causal effects in microbiomes using interventional calculus. Sci. Rep., 11, 5724.

Scutari, M. (2009) Structure variability in Bayesian networks. Available in: https://doi.org/10.48550/arXiv.0909.1685.

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. J. Stat. Softw., 35, 1-22.

Selman, B. and Gomes, C.P. (2006) Hill-climbing Search. In: Encyclopedia of Cognitive Science. Nadel, L. (Ed.), Wiley, New York.

Solé-Jiménez, P. et al (2021). Reshaping of gut microbiota in gilthead sea bream fed microbial and processed animal proteins as the main dietary protein source. Front. Mar. Sci., 8, 842.

Terova, G. et al. (2022). Highlights from gut microbiota survey in farmed fish - European sea bass and gilthead sea bream case studies. Aquaculture Europe, 47, 5–10.

Textor, J. et al. (2016) Robust causal inference using directed acyclic graphs: the R package 'dagitty'. Int. J. Epidemiol., 45, 1887-1894.

Vaidyanathan, R. et al. (2022) htmlwidgets: HTML Widgets for R. R package version 1.6.0. https://cran.r-project.org/web/packages/htmlwidgets/index.html.

Wickham, H. (2022) stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.5.0. https://cran.r-project.org/web/packages/stringr/index.html.

Yihui, X. et al. (2022) DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 0.26 https://cran.r-project.org/web/packages/DT/index.html.

Yuniarti, I. et al. (2021) An application of Bayesian Belief Networks to assess management scenarios for aquaculture in a complex tropical lake system in Indonesia. PLoS One, 16, e0250365.

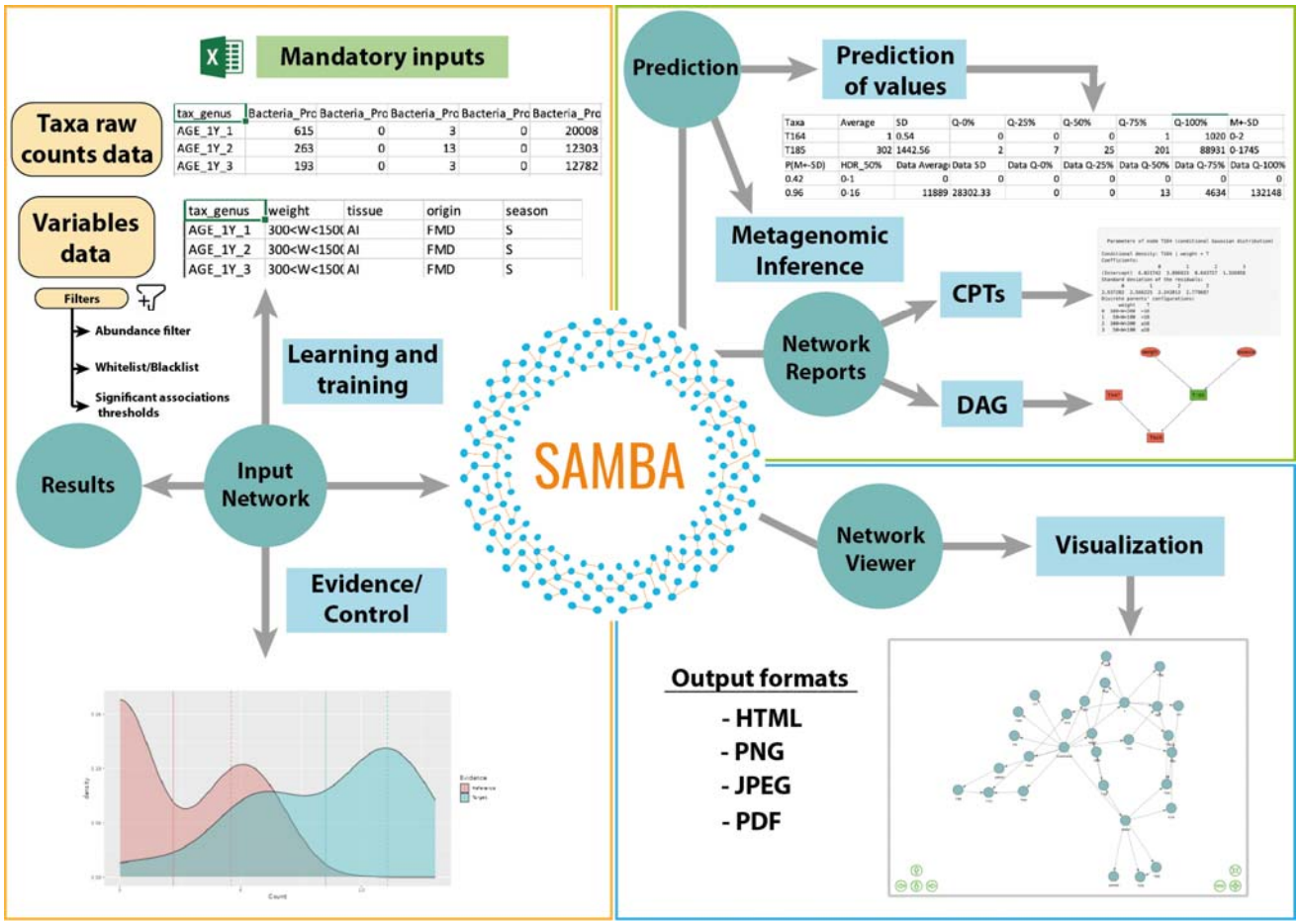
Zeileis, A. et al. (2008) Regression Models for Count Data in R. J. Stat. Softw., 27, 1-25.

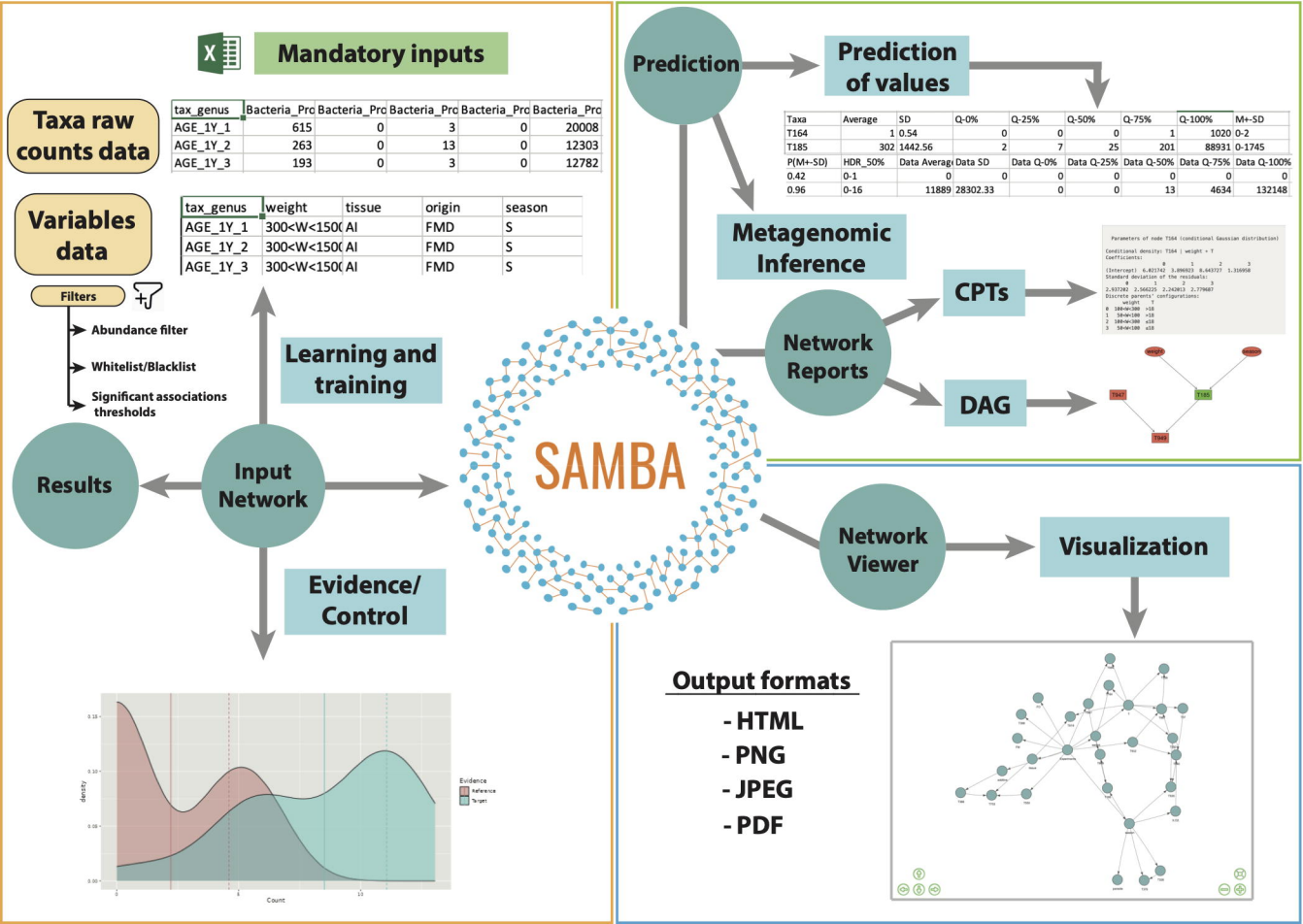
Figure Legends.

Figure 1.- SAMBA is implemented in 5 modules: 1) Input Network. A BN model can be created from the user input data using the Learning and training function. Also, values distribution for a condition or a combination of conditions is shown in the Evidence/Control interface. 2) Prediction. Prediction of taxa values and metagenome inference can be performed in this module. 3) Network Reports. Module to get Conditional Probability Tables (CPT) and/or Direct Acyclic Graphs (DAG) for a node. 4) Network Viewer. Module to visualize, edit, customize, and export networks. 5) Results. Module to download a zip file containing the Learning and training interface results.

Figures

Figure 1




Mandatory inputs
Taxa raw counts data
tax_genus | Bacteria_Pro | Bacteria_Pro | Bacteria_Pro | Bacteria_Pro | Bacteria_Pro
AGE_1Y_1 | 615 | 0 | 3 | 0 | 20008
AGE_1Y_2 | 263 | 0 | 13 | 0 | 12303
AGE_1Y_3 | 193 | 0 | 3 | 0 | 12782
Variables data
tax_genus | weight | tissue | origin | season
AGE_1Y_1 | 300<W<150(| AI | FMD | S
AGE_1Y_2 | 300<W<150(| AI | FMD | S
AGE_1Y_3 | 300<W<150(| AI | FMD | S
Filters
Abundance filter
Whitelist/Blacklist
Significant associations thresholds
Learning and training
Results
Input Network
Evidence/ Control
SAMBA
Prediction
Prediction of values
Taxa | Average | SD | Q-0% | Q-25% | Q-50% | Q-75% | Q-100% | M+-SD
T164 | 1 | 0.54 | 0 | 0 | 0 | 1 | 1020 | 0-2
T185 | 302 | 1442.56 | 2 | 7 | 25 | 201 | 88931 | 0-1745
P(M+-SD) | HDR_50% | Data Averag | Data SD | Data Q-0% | Data Q-25% | Data Q-50% | Data Q-75% | Data Q-100%
0.42 | 0-1 | 0 | 0 | 0 | 0 | 0 | 0 | 0
0.96 | 0-16 | 11889 | 28302.33 | 0 | 0 | 13 | 4634 | 132148
Metagenomic Inference
Network Reports
CPTs
DAG
Network Viewer
Visualization
Output formats
- HTML
- PNG
- JPEG
- PDF
Evidence
Reference
Target
density
Count