Fast and flexible analysis of linked microbiome data with mako

Lisa Röttjers and Karoline Faust

Mako is a software tool that converts microbiome data and networks into a graph database and visualizes query results, thus allowing users without programming knowledge to carry out network-based queries. Mako is accompanied by a database compiled from 60 microbiome studies that is easily extended with the user’s own data. We illustrate mako’s strengths by enumerating association partners linked to propanone production and comparing frequencies of different network motifs across habitat types.

Microbial association network construction and analysis have become well-established tools for the study of microbiomes. While meta-analytical approaches are increasingly adopted in a range of studies, no framework exists that provides flexible and intuitive methods for the storage of microbiome-derived networks. Network databases and query languages such as Neo4j and Cypher implement an alternative data storage and query model that has not been fully exploited for microbiome data yet. Here, we present mako (microbial associations catalog), which fills this gap, thereby enabling rapid and simple construction of network databases from microbiome data and metadata.

For macro-ecology, interaction databases such as Mangal provide a data specification that permits data sharing and reuse in R and Julia. String-DB fulfills a similar role for protein interaction databases. Such databases include validated and observed interactions between species or between proteins. Currently, no database storing associations between taxa has been published for microbiome data (although a static database with a web interface is under development by Hu et al., http://www.microbialnet.org/mind_home.html). However, a static network database does not accommodate the flexibility required for microbial associations, since inferred associations change depending on the experimental design used to generate the abundance data and the parameters of the tools used to infer the networks. Additionally, microbial network inference is a rapidly developing field and no standard protocols exist for the inference of these networks. Consequently, the analysis of microbial associations demands a more flexible and local setup that can integrate a wealth of data. So far, this requires a large amount of data processing and maintenance efforts.

Mako provides an interface between standard microbiome formats and Neo4j graph databases via a database schema based on semantic web ontologies (Fig. 1). This software package includes a range of methods to interact with Neo4j databases and the pattern-based query language Cypher, requiring only rudimentary computational skills. Additionally, mako includes a curated database derived from 60 separate data sets downloaded from QITA, a platform for hosting microbial studies that facilitates meta-analyses of this scale.

Network databases such as Neo4j differ from relational databases in that they do not store data as tables but as networks (graphs). While microbial association networks typically contain only links between different microbial taxa or abiotic factors, a graph database storing those association networks can contain different pieces of information in its graph as well. For example, a taxon’s phylum can be stored as a node in a graph database. The taxon’s phylum is then represented as an edge connecting the taxon identifier to the phylum. An edge is therefore a relationship containing two nodes—two pieces of information—and can have multiple meanings, depending on its label. In a graph database, an edge can be an association or an interaction, but it may also be the observation linking a taxon to a sample. Consequently, graph database representations of biological data look more like their conceptual representation. To illustrate, a taxonomic tree can be stored as a tree, a type of graph, rather than a table.

Most standard biological formats use tables of some form and therefore need to be processed to be compatible with graph databases. These processes—also referred to as extract, transform, load processes—can be combined with semantic web technologies such as ontology web language (OWL) to integrate and structure data from multiple sources. Ontologies, such as the National Cancer Institute (NCI) thesaurus that mako uses for its OWL file, are controlled vocabularies that are machine- and human-readable to improve the accessibility of biological data. Hence, mako includes a conceptual model written in OWL that can be used to check the integrity of the Neo4j database.

To illustrate the speed and ease of use of Cypher, the query language used to access Neo4j databases, we curated 60 BIOM files and generated 60 networks with FlashWeave, constructing a final Neo4j database that contains over 5 million relationships, connecting more than 100,000 nodes. This curated database is provided as a supplementary file with mako. On a Dell Precision laptop with a 3.00 GHz Intel Xeon computer processing unit (sixth generation), an M.2 SSD and 32 GB of RAM, uploading of all BIOM files and networks took approximately 12 min. Once this database was initialized, restoring the 1.322 GB database from a database dump took 11 s on the same machine.

A pattern-based language such as Cypher can be used to identify motifs. Network motifs may indicate the presence of specific dynamic behaviors. For instance, in microbial communities, a negative circuit implementing a rock-paper-scissors game can promote diversity. For association networks, many of these well-characterized motifs are not defined since they consist of directed relationships, while association networks are undirected. Other motifs have not been investigated extensively.

Cypher facilitates the investigation of such motifs in large data sets. To demonstrate this, we wrote Cypher queries to search for motifs that represent either three- or four-node subgraphs (Figure 2a). We used our 60-network database to explore certain combinations
of weights in maximally-connected subgraphs across four EMPO (Earth Microbiome Project Ontology) terms. Our results indicate that there are distinct differences between the four EMPO terms. First, there are more associations attributed to nonsaline and animal networks than could be expected, given that 33 and 15 of our 60 data sets had the Animal or Nonsaline EMPO terms, respectively. Second, animal-derived networks appear to have higher motif density. These results are supported by previous studies that found differences in edge density across biomes. However, our approach did not allow us to fully explore differences between plant and nonsaline microbiomes, since both studies may contain soil samples.

The motifs display another striking pattern: internally consistent motifs are overrepresented. Such motifs are balanced graphs, meaning that the product of their edge signs in every cycle is equal to one. In microbiological terms, the overrepresentation of balanced graphs means that in the presence of positive and negative edges, two microbes with a positive association between them are more likely to share negative relationships with other microbes.

In addition to assessing motif structure, taxonomic information can be included in queries. Formation of propionate via 1,2-propanediol can involve different species, depending on the substrate. Here, we looked at three different groups of substrates (sugars, lactate and deoxy sugars fucose and rhamnose), which can be converted to 1,2-propanediol via different genera (Figure 2b). We queried for associations between these and genera that convert 1,2-propanediol to propionate. We found several associations between genera that could theoretically produce propionate together from fucose or rhamnose, but only one association was found that could be linked to lactate degradation. For sugars, only Clostridium and Escherichia could be linked to propionate producers. Overall, the queries indicate that specific substrates could not be linked to associations with a single propionate producer.

To support the use of Neo4j databases for biological data, mako includes several modules. Specifically, the following features are supported:

1. Imports of biological data. Several types of biological data, commonly used in the analysis of association networks, can be imported via a simple application programming interface (API), command line interface or GUI. These include BIOM files, tab-delimited files, network files and edge lists of custom properties. These data are stored according to a clearly documented database schema, which supports development of custom Cypher queries. The mako software supports batch uploading of node data and uses indices to rapidly speed up database access operations.

2. Meta-analytical approaches. While the full scope of meta-analytical approaches is too broad to be included in this software package, several methods have been included that are potentially useful. These include options to make ‘overlay’ networks from combinations of networks, such as intersections, unions or differences. The software also includes methods for...
merging networks by taxonomic labels, so structure at higher taxonomic levels can be investigated.

(3) Integration with external software, including Cytoscape. The Neo4j database is one large graph that includes both metadata and network data. Loading this entire database into Cytoscape is not feasible, not only because the number of nodes in the Neo4j database can far exceed the number of nodes that can be rendered with Cytoscape, but also because the data model for Cytoscape includes fundamentally different representations of node and edge metadata. The mako software can extract more common network representations from the Neo4j database and import these directly into Cytoscape via an HTTP request, or supply these to other network analysis tools.

(4) Flexible query design. The database schema used by this software has been designed to accommodate a range of biological data. Any data that can be rewritten to an edge list (for example, property-taxon or property-edge) can be imported into the database. The online manual at https://ramellose.github.io/mako_docs/ contains a brief introduction to Neo4j database schema design, a detailed explanation of the database schema used by the software and some simple explanations of Cypher queries. Additionally, all scripts and queries used for examples shown in this paper have been provided as supplementary files to demonstrate flexibility of the Cypher query language for biological data.

The mako toolbox does not fully exploit the potential of graph databases. For instance, the database schema could be extended to ease standardized queries for different types of data. Especially for large omics data, the increase in speed that could result from additional standardization is likely to be substantial. Moreover, the analysis of such linked data in microbiome research is not well established and may require novel analytical methods. In conclusion, while still leaving room for improvement, the mako toolbox makes an important contribution to the adoption of graph databases in microbiome research.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-021-01335-9.

Received: 19 February 2021; Accepted: 1 November 2021; Published online: 9 December 2021

References
1. Röttjers, L. & Faust, K. From hairballs to hypotheses—biological insights from microbial networks. FEMS Microbiol. Rev. 42, 761–780 (2018).
2. Jackson, M. A. et al. Detection of stable community structures within gut microbiota co-occurrence networks from different human populations. PeerJ 6, e4303 (2018).
3. Wang, H. et al. Combined use of network inference tools identifies ecologically meaningful bacterial associations in a paddy soil. Soil Biol. Biochem. 105, 227–235 (2017).
4. Poisot, T. et al. mangal–making ecological network analysis simple. Ecography 39, 384–390 (2016).
5. Szklarczyk, D. et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 47, D607–D613 (2019).
6. Gonzalez, A. et al. QIITA: rapid, web-enabled microbiome meta-analysis. Nat. Methods 15, 796–798 (2018).
7. Miller, J. J. Graph database applications and concepts with neo4j. In Proc. Southern Association for Information Systems Conference, Atlanta, GA, USA Vol. 2324 (ed. Fitzpatrick, L.) 141–147 (AIS, 2013).
8. Bansal, S. K. Towards a semantic extract-transform-load (ETL) framework for big data integration. In Proc. 2014 IEEE International Congress on Big Data (eds Chen, P. & Jain, H.) 522–529 (IEEE, 2014).
9. Noy, N. F. et al. Creating semantic web contents with protege-2000. IEEE Intell. Syst. 16, 60–71 (2001).
10. Kerr, B., Riley, M. A., Feldman, M. W. & Bohannan, B. J. Local dispersal promotes biodiversity in a real-life game of rock–paper–scissors. Nature 418, 171–174 (2002).
11. Ma, Z. S. & Ye, D. Trios-promising in silico biomarkers for differentiating the effect of disease on the human microbiome network. Sci. Rep. 7, 13259 (2017).
12. Thompson, L. R. et al. A communal catalogue reveals earth’s multiscale microbial diversity. Nature 551, 457–463 (2017).
13. Ma, B. et al. Earth microbial co-occurrence network reveals interconnection pattern across microbiomes. Microbiome 8, 82 (2020).
14. Faust, K. et al. Cross-biome comparison of microbial association networks. Front. Microbiol. 6, 1200 (2015).
15. Louis, P. & Flint, H. J. Formation of propionate and butyrate by the human colonic microbiota. Environ. Microbiol. 19, 29–41 (2017).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021, corrected publication 2022
Methods

**QIITA-derived networks included with mako.** To demonstrate the Neo4j database schema and mako's functionality, we used 60 files that were downloaded from QIITA. Processed files and derived microbial association networks have been included as part of the database dump.

**Processing 60 data sets from QIITA.** We downloaded 60 files that could roughly be assigned to the following three categories: human gut microbiome, soil microbiome (including terrestrial plants) and aquatic microbiome. Of the 60 files, 37 had an empo_2 column containing EMPO level 2 annotations. For each of these, the most common empo_2 annotation was used to label the network. The remainder had empo_2 annotations added manually; these were either stored in a differently named column or derived from the study description. After manual curation, 33 data sets had the EMPO 2 label 'Animal'; 15 were labeled 'Nonsaline'; eight were labeled 'Plant' and four were labeled 'Saline'. Given the richness of metadata available in QIITA, this data was curated first to speed up database operations. Values that were identical across a data set were filtered. We also chose to remove values that we considered less relevant for future analysis; these values included the barcode sequence, the collection timestamp, the run date and the run prefix.

To omit statistical issues with missing values, columns with more than three missing values were also filtered. For one specific data set, columns with only 'missing', 'not applicable' and 'not collected' entries were removed. Finally, we found that columns with mostly numeric values contained characters referring to missing samples; such columns were also removed if they had more than five values containing characters.

For a full overview of all downloaded files, we refer to Supplementary Table 1.

**Running network inference with FlashWeave.** Taxon abundances were first collapsed to the genus level identifier; only taxa that were present in >20% of samples were retained. Taxa that fell below the prevalence threshold were merged into a single synthetic taxon so that the total sum of abundances was not disturbed. The resulting abundance tables were used to infer networks with FlashWeave v.0.18.0, Julia v.1.5.3 (ref. 3). FlashWeave was set to sensitive, nonheterogeneous and to only consider associations with more than ten observations. Since FlashWeave incorporates a clr transformation, it handles compositional data appropriately without additional preprocessing steps.

**Motif analysis with Neo4j.** In a Python v.3.6.3 environment, we used the mako software package in combination with pandas v.1.1.5 to generate tables of motif counts. The Python script used to query motifs is available via the mako Zenodo software package in combination with pandas v.1.1.5 to generate tables of motif counts. The Python script used to query motifs is available via the mako Zenodo software package.

**Creating a database dump with Neo4j.** The mako software was used to upload processed BIOM files and network files to a local Neo4j database. Next, the neo4j-admin script was used to create a dump of the complete database. This file, called mako.dump, has also been made available through Zenodo. The file can be used to rapidly restore the complete database.

**Database schema.** A core component of mako is a database schema constructed from several OWL terms. This database schema also defines constraints for the data and provides guidance on how data need to be uploaded to become easily accessible. Most functionality of the API uses the schema to construct queries for uploading BIOM files, checking edge weights or finding associations between related species.

To properly integrate with Neo4j’s ability to work with OWL, each node and relationship label is derived from an existing ontology, the NCI Thesaurus. However, since relationships are not well-defined across biological ontologies, constraints for relationships were defined manually with Protégé. For each relationship, domains and ranges (target and source nodes) were specified. While these axioms are not added to Neo4j as actual constraints, mako can run checks to see whether domain and range axioms are violated by any nodes or relationships in the database.

The complete database schema represents associations as separate nodes, rather than a relationship between two taxa. This makes it easier to query association properties. This database schema represents a tradeoff between both flexibility and speed since the Cypher query planner needs to scan fewer nodes if the labels are provided. Consequently, separate labels are used to indicate pieces of information common to all count tables and microbial association networks. Neo4j offers the opportunity for advanced users to extend this database schema with new labels, so they can write faster queries for specific metadata.

**Interfacing with the Neo4j database.** The mako software was developed in Python v.3.6.6 and uses the Python API provided by Neo4j (ref. 4). Simply put, the software converts each of the input files (BIOM, tab-delimited files, network files, edge lists) into a list of dictionaries. Each dictionary contains the necessary information, for example, node names or other properties, that needs to be provided to parameterized queries. Rather than running individual queries, all larger queries are run as batch queries so that the Cypher query planner can reuse the query plan each time.

Since node labels cannot be added as query parameters, separate queries need to be written for each unique node label. In some cases, for example, taxonomies, queries are constructed dynamically in Python and then used to construct separate batch queries for each taxonomy node. This permits reuse of query designs across node labels.

**Manipulating networks in Neo4j and Python.** Some utilities are available for meaningfully interacting with the Neo4j database. For example, mako queries edges that are present in multiple networks and uses this to construct ensemble networks (for example, intersection, union or difference operations). The software can also find pairs of edges, that is, edges that have identical taxonomic labels at both ends and merge these. Additionally, mako includes functionality for exporting networks to Cytoscape. While a Cytoscape Neo4j plugin is available, this does not accommodate translations from one database schema to another. The Neo4j database schema used by mako was specifically designed to support meta-analyses of large amounts of information and therefore supports databases with millions of nodes. In contrast, Cytoscape renders networks (partially) and may therefore not be able to represent a full database stored in Neo4j. For this reason, mako includes scripts that extract (specific) networks from the Neo4j database, convert them to a Cytoscape-compatible json file and use the Python requests HTTP package to transmit these json files to a running instance of Cytoscape.

One potential issue with Cypher is that patterns can be matched in several different ways at the same time the query planner comes across them; the order of patterns is returned, even if they were already part of the same pattern but in a different order. Consequently, unlike SQL, some extra parsing of the results can be required. For example, the pattern (a)–(b)–(c)–(a) can be matched to (Node 1)–(Node 2)–(Node 3)–(Node 1); but also to (Node 2)–(Node 3)–(Node 1)–(Node 2). Where applicable, such query parsing is implemented in mako. Example functions to collect unique patterns have also been included in the supplementary script available through Zenodo.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

The complete Neo4j database used in this paper has been used to generate a Neo4j dump file. An archived version of this file has been submitted to Zenodo. Instructions for reconstructing this database are available in mako’s documentation. All BIOM files were downloaded from QIITA and a full overview of these studies can be found in Supplementary Table 1.

**Code availability**

The latest version of the mako software can be downloaded via https://github.com/ramellose/mako/. An archived version, including the script used to quantify motifs in the Neo4j database, has been submitted to Zenodo. All mako source code and the included script are available under the Apache License 2.0. An extensive manual for running mako and Neo4j (via Docker) is available via https://ramellose.github.io/mako_docs/. Additionally, an accompanying compute capsule allows a demo of mako and a Neo4j database to be run without requiring any installations.

**References**

16. Tackmann, J., Rodrigues, J. F. M. & von Mering, C. Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. Cell Syst. 9, 286–296 (2019).

17. Röttjers, L. & Faust, K. Fast and flexible analysis of linked microbiome data with mako. Zenodo https://doi.org/10.5281/zenodo.4946425 (2021).

18. Conway, J. R., Lex, A. & Gehlenborg, N. Upsetr: an R package for the visualization of intersecting sets and their properties. Bioinformatics 33, 2938–2940 (2017).

19. Soutos, N. et al. Nci thesaurus: a semantic model integrating cancer-related clinical and molecular information. J. Biomed. Inform. 40, 30–43 (2007).

20. Sumner, G. et al. cynos4: connecting neo4j and cytoscape. Bioinformatics 31, 3868–3869 (2015).

21. Röttjers, L. & Faust, K. Fast and flexible analysis of linked microbiome data with mako. Code Ocean https://doi.org/10.24433/CO.0482418.v1 (2021).

**Acknowledgements**

This project was supported by the KU Leuven under grant no. STG/16/006. K.F. has received funding from the European Research Council under the European Union’s Horizon 2020 research and innovation program under grant agreement no. 801747.

**Author contributions**

L.R. developed the software, carried out the case studies and drafted the paper. K.F. contributed to the design of the software and the case studies, and provided substantial revisions of the paper. Both authors read and approved the final paper.
Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-021-01335-9.

Correspondence and requests for materials should be addressed to Karoline Faust.

Peer review information Nature Methods thanks Jianguo Xia and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available. Lin Tang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
  - Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  - Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection No software was used for data collection.

Data analysis Networks were inferred with FlashWeave version 0.18.0, Julia version 1.5.3. All data analysis was done in Python version 3.6.3 and R version 4.0.2. Pandas version 1.1.5 was used to process motif counts. Figures were generated with ggplot version 3.3.2, viridis version 0.5.1 and UpSetR version 1.4.0.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The complete Neo4j database used in this manuscript has been used to generate a Neo4j dump file. An archived version of this file has been submitted to Zenodo (https://doi.org/10.5281/zenodo.4946415). Instructions for reconstructing this database are available in mako's documentation. All BIOM files were downloaded from Qiita and a full overview of these studies can be found in Supplementary Figure 1.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences    ☐ Behavioural & social sciences    ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | 60 independent microbial abundance data sets were included to showcase the software described in the manuscript. Since this is an exploratory study and no statistical tests were carried out, no sample size calculations were performed. |
| Data exclusions | No data sets were omitted from analyses. |
| Replication | Since the presented data is a case study to showcase software, replication is not applicable. |
| Randomization | This manuscript does not include any samples that had to be randomized. |
| Blinding | No blinding was carried out for this study as this was an exploratory analysis. |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems | n/a | Involved in the study |
|----------------------------------|-----|-----------------------|
| Antibodies                      | ☐   | ☒                     |
| Eukaryotic cell lines           | ☐   | ☒                     |
| Palaeontology and archaeology   | ☐   | ☒                     |
| Animals and other organisms     | ☒   | ☐                     |
| Human research participants     | ☒   | ☐                     |
| Clinical data                   | ☒   | ☐                     |
| Dual use research of concern    | ☒   | ☐                     |

| Methods                         | n/a | Involved in the study |
|---------------------------------|-----|-----------------------|
| ChIP-seq                        | ☒   | ☐                     |
| Flow cytometry                  | ☒   | ☐                     |
| MRI-based neuroimaging          | ☒   | ☐                     |