SuperMoM: A graph-based database to store and analyze microbial community omics and environmental data

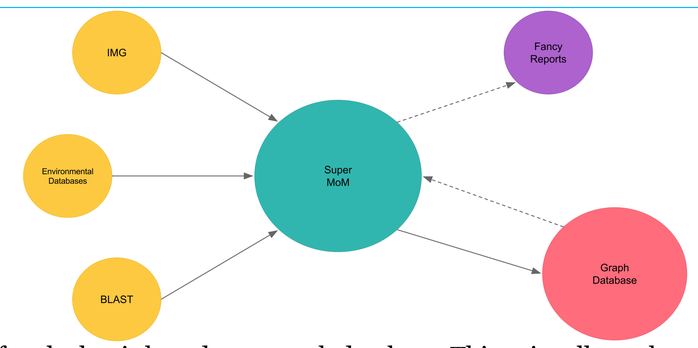
Sunit Jain, Gregory Dick

Dept. Of Earth & Environmental Sciences, University of Michigan

Abstract

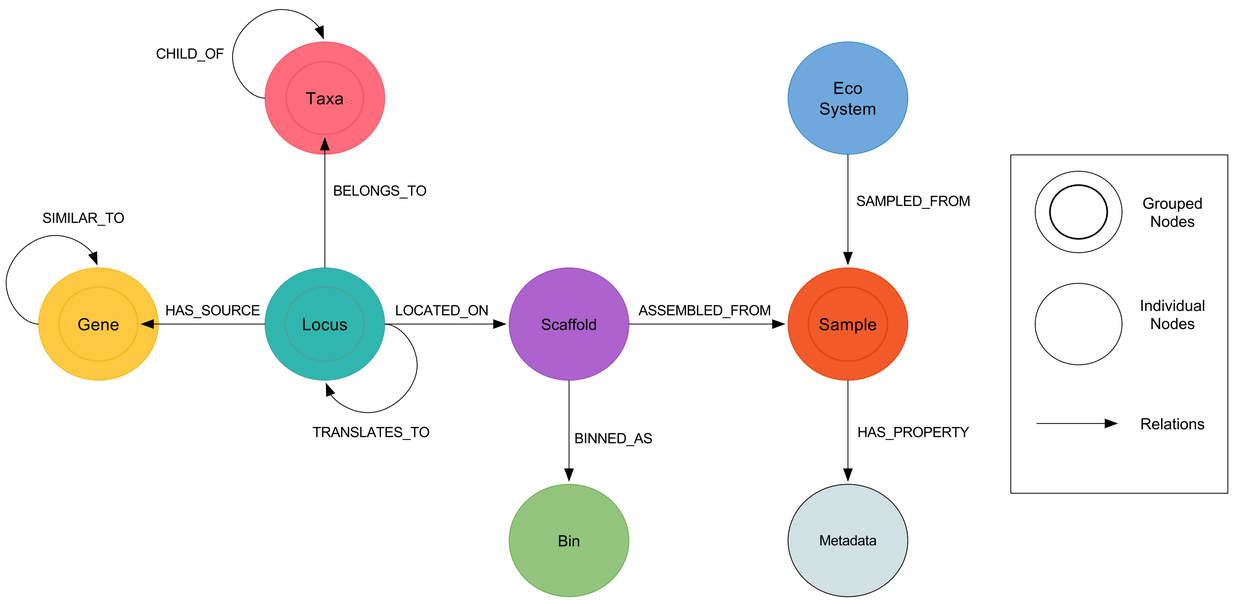
Rapidly advancing DNA sequencing technologies provide new opportunities for studying cyanobacterial harmful algal blooms (CHABs). Specifically, we can now generate enormous datasets on the composition of genomes, transcriptomes, and proteomes of microorganisms as they occur in natural microbial communities, thus providing unprecedented insights into the ecology of CHABs. However, processing, storing, and analyzing the massive datasets now routinely produced presents new challenges. Substantial effort has been invested into infrastructure for archiving and analyzing sequence data, leading to large data warehouses like Genbank[1] and IMG[2,3]. In addition, smaller, more specialized databases such as ProPortal[4] have been developed to focus on specific organisms like Prochlorococcus. A major hurdle in mining such information is that the data is often large, multidimensional, encoded in ontologies and in a wide variety of formats. As these databases grow they become more unwieldy. For relational databases, complex relational queries may have unacceptably long compute times. To address these issues we present SuperMoM (**S**uper **M**eta**o**mics **M**iner), a flexible and scalable graph database. Due to the graph nature of the database, it is optimized for handling complex relational queries without adversely affecting performance. We believe this database will facilitate exploration and comparison of multidimensional sequence data in the context of environmental data, thus allowing users to address more complex scientific questions. This tool is under active development and alpha releases can be downloaded from its github page: https://github.com/sunitj/SuperMoM.

SuperMoM



SuperMoM is a suite of tools that is based on a graph database. This suite allows the user to integrate data from different disciplines into a single database. Since it is based on a graph structure, extracting results is as simple as traversing from one node to another. Extracted data can be readily pulled into any number of scripting languages to make customized reports and visualizations. SuperMoM also employs a variety of algorithms written in Perl and R to make extraction and visualization easy and interactive.

Database Schema



Here is a simplified conceptual diagram of the biological portion of the database. A graph database comprises of three important components, “Nodes”, “Relations” and “Properties”. Note a fourth component, “Grouped Nodes”, which are a collection of similar nodes grouped together for simplicity. A relationship defines how two nodes are connected. All nodes and relations can have properties specific to them. A query in such a database is simply a pattern of nodes and relations constrained by their properties. It is this property-graph model that allows us to search for intricate patterns in our data that we would have otherwise missed.

Properties

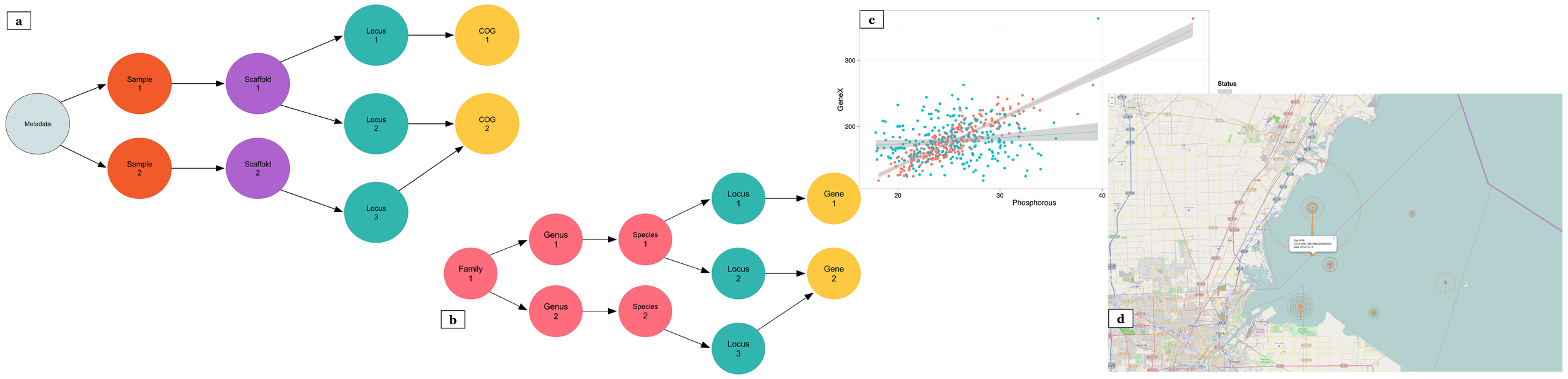
Node	Subtype	Property	Example
Ecosystem	Ecosystem	Category	Aquatic
		Type	Freshwater
		Subtype	Limnetic
		Specific	Lake Erie
Metadata	Metadata	Latitude	45.062
		Longitude	-83.431
		Turbidity	23
		CHLa Concentration	466
		Etc.	
Scaffold	Scaffold	Length	400,000
		GC	58%
Bin	Bin	Name	Microcystis
		Confidence	97%
Sample	Metagenomic	ID	Sample ID
	Meta Transcriptomic	Name	CHABs_1
	MetaProteomic	Site	Site_1
Locus	Genomic	Name	Locus_1
	Proteomic	Coverage	
Gene	COGs	ID	COG0438
	Pfam	Name	Glycosyl transferase
	KEGG	Function	Cell wall biosynthesis
Taxa	Species	Taxon ID	449447
	Genus	Genetic Code	11
	Domain	Kmer Signature	[Vector]
	Phylum	Average GC	58%
	Class	Strain	NIES-843
	Order	Species Name	Microcystis *aeruginosa*
	Family	Genus Name	Microcystis

From	Relation	To	Property	Example
Sample	SAMPLED_FROM	Ecosystem	Year	2015
			Month	April
			Date	13
			Time	1830
	HAS_PROPERTY	Metadata	N/A	N/A
Scaffold	ASSEMBLED_FROM	Sample	Assembler	IDBA-UD
			Command	./idba_ud ...
			Reads_Mapped	65%
	BINNED_AS	Bin	Method	ESOM
			Confidence	100%
Locus	LOCATED_ON	Scaffold	Start	4000
			End	6000
			Strand	+
	BELONGS_TO	Taxa	Confidence	97%
	HAS_SOURCE	Gene	Identity	73%
	TRANSLATES_TO	Locus	N/A	N/A
Gene	SIMILAR_TO	Gene	Identity	97%
			Method	BLASTN
Taxa	CHILD_OF	Taxa	N/A	N/A

Table2 (top) shows the Relationship types between Nodes and a hypothetical example for the properties in each relation. The colors in the ‘from’ and ‘to’ columns correspond to node colors in the schema.

Table1 (left) depicts a Node, it’s subtype, properties and a hypothetical example. The colors in the node column correspond to node colors in the schema.

Results



Querying the database is as easy as drawing the pattern of relations on a white board. Shown here in (**a**) is a conceptual diagram of a sample query that will search the database and plot the expression of a gene of interest as a function of environmental conditions. For example a user might wish to plot omics data as a function of various environmental variables such as temperature or the concentration of a nutrient in a system. Similarly, (**b**) is a conceptual diagram to analyze “pangenomes” of organisms that belong to the same taxonomic group (family in this case). Such a framework could be used to efficiently compare the gene content of related organisms, or the taxonomic distribution of specific genes.

In order to ask these important questions a user needs to search the database. Searches can be executed in one of three ways:

1. Through the Neo4j[5] web browser interface.
2. From within SuperMoM; or
3. Through the user’s choice of programming language that supports REST.

Using SuperMoM, the user will be able extract this information from the database and visualize it using any number of methods, from simple scatter plots (**c**) to more interactive map based plots (**d**). The database is open-sourced and uses the REST-API for queries. The user will need the latest versions of BioPerl[7], R and Neo4j installed to make use of this tool.

Future Work

We are continuously developing more algorithms and adding new data collected from CHABs every day. Our aim would be to eventually expand this database to include additional data intensive bio-geo-chemical projects. Some additions that you can expect to see in the coming months are:

- Support for integration with additional open source works like the Bio4j[6] project.
- Adding more graph traversal and network analysis algorithms.
- Adding support for metaproteomic data.
- Adding documentation and tutorials.

Conclusions

An intuitive way of searching for connections makes this database an immensely useful tool. With SuperMoM, we hope to provide a platform for researchers to integrate, explore and visualize the hidden patterns that emerge when a variety of interdisciplinary data is meaningfully brought together.

As of now, we have tested this database with three metagenomic datasets and the results have been promising. This project remains under active development and is regularly updated at https://github.com/sunitj/SuperMoM For questions regarding the project please contact *Sunit Jain* at sunitj@umich.edu.

References

1. Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. Nucleic Acids Research, 41(Database issue), D36–42. doi:10.1093/nar/gks1195
2. Markowitz, V. M., Chen, I.-M. A., Chu, K., Szeto, E., Palaniappan, K., Pillay, M., ... Kyrpides, N. C. (2014). IMG/M 4 version of the integrated metagenome comparative analysis system. Nucleic Acids Research, 42(Database issue), D568–73. doi:10.1093/nar/gkt919
3. Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Pillay, M., ... Kyrpides, N. C. (2014). IMG 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Research, 42(Database issue), D560–7. doi:10.1093/nar/gkt963
4. Kelly, L., Huang, K. H., Ding, H., & Chisholm, S. W. (2012). ProPortal: a resource for integrated systems biology of Prochlorococcus and its phage. Nucleic Acids Research, 40(Database issue), D632–40. doi:10.1093/nar/gkr1022
5. Neo4j an open-source graph database, implemented in Java (https://neo4j.com)
6. Pareja-tobes, P., Tobes, R., Manrique, M., Pareja, E., & Pareja-Tobes, E. (2015). Bio4j: a high-performance cloud-enabled graph-based data platform. bioRxiv. Cold Spring Harbor Labs Journals. doi:10.1101/016758
7. Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., ... Birney, E. (2002). The Bioperl toolkit: Perl modules for the life sciences. Genome Research, 12(10), 1611–8. doi:10.1101/gr.36160

Acknowledgments

This work was supported by a grant from the University of Michigan Water Center, which is supported by the Erb Family Foundation and the U-M Provost. Annotation for the sample data was provided by the Joint Genome Institute Integrated Microbial Genomes Expert Review with Metagenomic Samples. Thanks also to the members of Michigan Geomicrobiology Lab, the Microbial Evolutionary Ecology Lab and the Duhaime Lab for their feedback.