Biological Databases for Linking Large Microbial and Environmental Datasets

C. Jacob, A.D. Kent, B.J. Benson, R.J. Newton, and K.D. McMahon

Abstract— **New analytical techniques in microbiology have created the potential to investigate microbial communities, their interactions, and their role in ecosystem functions in novel ways. Combinations of such techniques allow the rapid generation of large datasets describing microbial community composition and variation across time and space. In order to address ecologically relevant questions, microbial community datasets must be linked with related environmental datasets. This challenging task is made feasible in a rich data mining environment through a complex data model and interactive querying tools against a relational database. This paper discusses the motivation and design for one such microbial ecology database at the North Temperate Lakes Microbial Observatory and the research questions made tractable by this informatics project. The design principles and data-model used during database development are presented. Architecture that supports the progressive evolution of the informatics system is also discussed. Interactions with the user community in data model development were essential. This application is custom-designed to the needs and objectives of linking microbial and environmental questions, highlighting contributions from informatics to ecology. The bio-data model reflects the data mining regimes of the microbial disciplines and supports research questions that could not have been asked without such informatics tools. The project also serves as an illustrative case study in the design of data models and information systems, not only for microbial-environmental datasets, but in a broader perspective, for other biological databases that could adapt the techniques used here for data integration and mining.**

Index Terms—**Biological Databases, Microbial and Environmental Datasets, Scientific Database Design.**

I. Introduction

The study of microbial populations in natural environments has been hindered by the inability to easily cultivate the vast majority of environmental microbes. To overcome this limitation, microbial ecologists have developed a variety of cultivation-independent approaches to access the diversity of prokaryotic populations. Such approaches vary in the quality of information they provide about microbial communities. Approaches involving sequence analysis of cloned DNA result in high quality data that can be used to define phylogenetic relationships among community members. This fine-scale phylogenetic resolution comes at the expense of reducing sample throughput and also sacrifices complete sampling coverage of the microbial community. At the other extreme, community "fingerprint" (DNA fragment analysis) approaches are commonly used to quickly assess the diversity of microbial communities. While these methods rapidly generate an overview of the microbial community and readily lend themselves to comparisons of community diversity and composition among many samples, taxonomic information about the microbial community is compromised. Choosing among the various cultivation-independent approaches requires a trade-off between phylogenetic resolution and sample throughput.

Many of the cultivation-independent methods in use by microbial ecologists are at least partially automated, allowing the rapid generation of very large datasets. Handling this data volume alone requires sophisticated data management; moreover, microbial ecology studies seek to place this information about microbial community composition and variability in the context of ecosystem function. In order to address ecologically relevant questions, these microbial community data must be linked to data describing environmental conditions and ecosystem processes. The informatics tools designed by the North Temperate Lakes Microbial Observatory seek to leverage the strengths of two cultivation-independent approaches for microbial community analysis, and connect these data with extensive long-term environmental datasets maintained by the North Temperate Lakes Long Term Ecological Research site. Our research goal is to analyze the dynamics of specific microbial populations in the context of environmental data in order to hypothesize about the relationship between microbial community structure and ecosystem processes that influence, and are influenced by, microbes.

II. Tools used & Configuration

Several standard software packages and programming languages available commercially are used to implement the system described in this paper. The database was implemented and maintained on a SUN Solaris Server [1] running Oracle

Cheryan Jacob, Research Assistant, University of Wisconsin-Madison (phone: 608-628-7455; fax: 608-890-0184; email: cjacob@wisc.edu)
Dr. Angela Kent, Research Associate, Center for Limnology and Civil and Environmental Engineering, University of Wisconsin-Madison
Dr. Barbara Benson, Associate Scientist, Center for Limnology, University of Wisconsin-Madison
Ryan Newton, Research Assistant, Microbiology Doctoral Training Program, and Civil and Environmental Engineering, University of Wisconsin-Madison
Prof. Katherine McMahon, Civil and Environmental Engineering and Microbiology Doctoral Training Program, University of Wisconsin-Madison

DBMS [2]. This database is maintained as a part of the NTL-LTER (North Temperate Lakes Long Term Ecological Research) Information Management System [3]. The server also runs Apache Web Servers [4] and Jakarta Tomcat Servers [5] with virtual hosts to manage and make available online various web interfaces for the scientists and public using various custom-made web schemas [6]. A three layer architecture using Oracle at the database or persistence layer, JAVA [7] at the Business Logic Layer and JSP [8] and Java Script at the User Interface Layer enables porting the data available in the database to the World Wide Web.

III. System Architecture

The system architecture is comprised of a 3-layered architecture custom designed to be adaptive to the requirements of a changing bio-informatics regime. The architecture aims to be modular enough so that each of the layers could be replaced with another technology while seamlessly interacting with each other to provide optimal levels of service to the scientists and other users accessing the large amounts of data available to them through the system. The different system layers are explained in detail below.

A. Persistence Layer:

The persistence layer is designed using relational database principles, normalized as far as possible given the nature of methods involved in data collection and implemented with extensive discussion between microbial scientists and information managers.

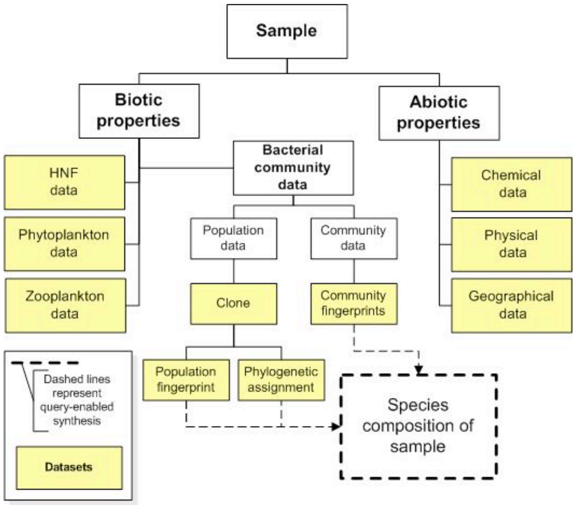


Figure 1: Persistence layer data model.

Figure 1 illustrates a simplified version of the data model that is implemented in the database/persistence layer. Environmental (abiotic) data from the different lakes are stored in the data tables containing chemical, physical and geographical parameters pertaining to each lake, keyed by sample date and lake code. The biotic data generated from each sample include species and abundance data for various planktonic populations. Visual identification and enumeration is possible for heterotrophic nanoflagellates (HNF), phytoplankton and zooplankton populations. The bacterial community data represents a special case where the determining the species composition involves a synthesis between two sets of data generated using molecular microbial ecology tools. For ease of discussion here, we will distinguish between “communities” and “populations”. Communities are comprised of many populations, and we can sample either the community or the population, the latter being more labor intensive and costly. The community “fingerprint” data represent the high-throughput but low phylogenetic resolution data generated from individual samples. The low-throughput data are derived from bacterial community DNA from many samples pooled for the analysis of specific bacterial populations present in a set of samples. Sequence data generated from the cloned DNA derived from multiple samples is used to determine the phylogenetic affiliation of specific microbial populations after comparison with public databases such as GenBank and the Ribosomal Database Project (RDP) [10,11]. Even higher quality phylogenetic data may be produced manually using sequence analysis tools such as ARB [12].

The labor and expense required to sample populations (especially developing clone libraries and generating phylogenetic assignments) make it desirable to extend the utility of these data beyond the characterization of the species present in an individual sample. To accomplish this extension, the unique molecular signature of individual microbial populations (population “fingerprint”) is also determined as part of the analysis of cloned DNA. The community fingerprints represent combinations of such signatures produced by the populations present in a given sample. By linking our library of phylogenetic information about specific microbial populations and their molecular fingerprints to the community fingerprint of a particular sample (Figure 2), we can rapidly generate a census of any microbial community.

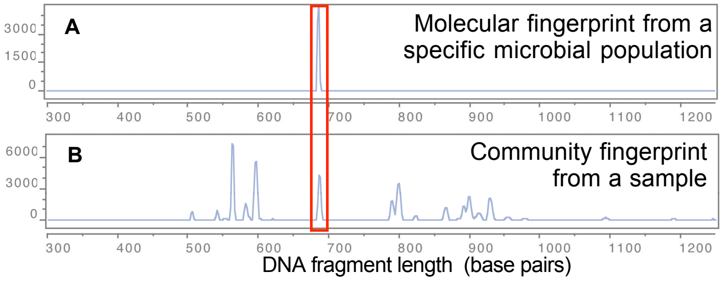


Figure 2: Characterizing the species composition of a sample involves matching the DNA fingerprint of a known microbial population (A) with the DNA fingerprint derived from the microbial community present in a sample (B). This matching process employs size tolerances (red) empirically determined for individual DNA fragments.

This approach maximizes the use of the high-quality bacterial population data and adds value to the lower-resolution (but high-throughput) bacterial community data. The North Temperate Lakes Microbial Observatory alone has generated more than 1400 community fingerprints from individual samples, and over 4000 DNA sequences and phylogenetic assignments derived from cloned DNA. The magnitude of these datasets prohibits manual synthesis of the species composition of all samples in this manner, especially given the constraint of filtering both community and population "fingerprints" through user-defined size tolerances for population-specific DNA fragments (Figure 2). In our database system these size tolerances are also stored in a data table and employed during the dynamic matching process that allows users to generate a list of microbial populations present in any given sample. Because both biotic and abiotic data are keyed by lake and sample identifiers (like sample data, sample depth etc.), specific information about microbial populations can be associated with environmental parameters of interest.

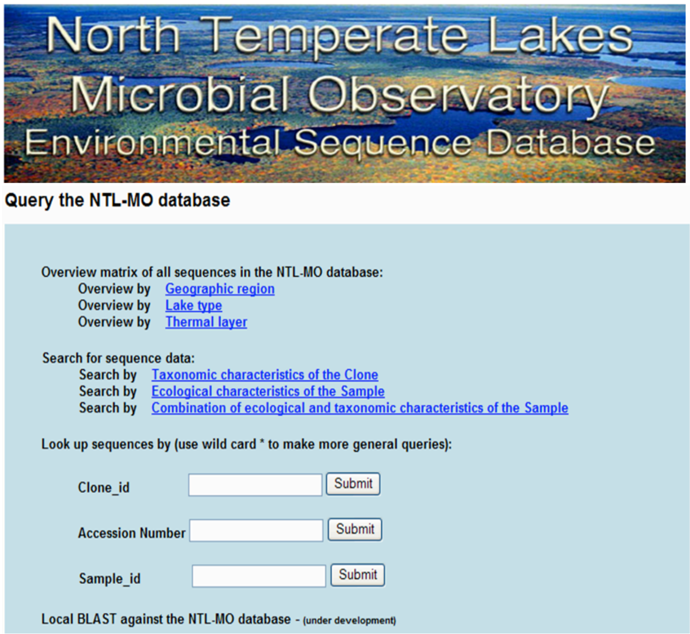
B. *Business Logic Layer:*

The business logic layer is implemented in Java and conforms to the NTL-LTER (North Temperate Lake Long Term Ecological Research) Information Management system for standardizing the management of various environmental, limnological and microbiological datasets collected over the years. This middle layer is responsible for transporting information and datasets between the persistence layer and the user interface layer. The queries made at the user interface are processed into SQL statements that the user interface layer sends to the business logic layer via custom made Java classes. The business logic layer opens JDBC connections to the persistence layer and retrieves result sets from this layer. The result sets thus retrieved are processed and formatted for display in the user interface layer.

The customizable web database applications implemented at the NTL-LTER (North Temperate Lake Long Term Ecological Research) [3] website also use this layer to intelligently read metadata stored in the database layer and dynamically create the parameters for the user interface layer based on queries made by the web user on the user interface layer [6].

C. *User Interface Layer:*

The user interface layer is implemented in JSP.. This layer has extensively used the web database application [6] designed by NTL-LTER-IM (North Temperate Lake Long Term Ecological Research- Information Management) to port data dynamically to the web. Several additional custom functions specific to microbial datasets and linking them to environmental datasets were created, and the existing functionality was enhanced to permit maximum standardization with the existing system[6] while making the web interface conform to the requirements of the microbial researchers (as shown in Figure 3). The web interface and available queries, as well as its structure and functionality can be seen at http://microbes.limnology.wisc.edu



North Temperate Lakes
Microbial Observatory
Environmental Sequence Database

Query the NTL-MO database

Overview matrix of all sequences in the NTL-MO database:
Overview by Geographic region
Overview by Lake type
Overview by Thermal layer

Search for sequence data:
Search by Taxonomic characteristics of the Clone
Search by Ecological characteristics of the Sample
Search by Combination of ecological and taxonomic characteristics of the Sample

Look up sequences by (use wild card * to make more general queries):

Clone_id
Accession Number
Sample_id

Local BLAST against the NTL-MO database - (under development)

Figure 3: Portal for the user interface layer.

The particular structure of the database enables microbiologists and other researchers to make complex, layered queries to get results and insights dynamically. An example of such a complex query is selecting all the bacteria from a particular lake type given specified environmental parameters or ranges of parameters along with a given set of microbiological parameters or their ranges and linking the two together. This type of query was a prohibitively complex problem that could not be done using traditional methods of data analyses and storage. The ability to accomplish this type of analysis clearly illustrates the advantages of having such a microbial-environmental database and the software for dynamic querying. The technology has the potential to change the way scientists in this field look at data and methods for analysis.

Bacterial populations
Gammaproteobacteria
Actinobacteria
Firmicutes
Deltaproteobacteria
Betaproteobacteria
Bacteroidetes
Alphaproteobacteria
Verrucomicrobia

Table 1. Bacterial phyla detected in a single sample

IV. Results

A. *Simple queries*

The composition of the microbial community detected in a single sample is easily determined by cross-referencing the community fingerprint of a sample with the library of population-specific DNA fingerprints. The bacterial populations detected in a humic lake on 27 July 2000 are shown in Table 1. This basic query has been programmed into the user interface layer and may be executed by simply entering a sample identifier. Other simple queries available through the user interface are queries on individual records of cloned DNA fragments. Entry of a clone identifier calls all information relevant to a specific bacterial population, including the DNA sequence, phylogenetic assignment for that sequence, and the population-specific fingerprint. Additional queries are automatically executed when the clone records are accessed. These compare the population fingerprint of a specific clone to all community fingerprints in the bacterial community dataset, and display a list of samples where this population was detected as part of the clone record.

B. *Complex queries*

The simple queries described above are executed on the bacterial community data alone. To add a more ecological perspective to microbial community data, we must consider the interactions of bacterial populations with other organisms, as well as feedbacks between these populations and environmental factors. Using the chemical and physical datasets associated with each lake or sample, we may constrain a query to list only the bacterial populations present in a specific niche. For example, we can mine our dataset for instances where potential disease-causing bacteria were detected in clear lakes in northern Wisconsin during the summer months of 2002. Such a query cross-references the bacterial community and population datasets with chemical, physical and geographical datasets. Similarly, we can examine the range of water temperature or pH or nutrients where specific populations of bacteria (or other organisms) are detected.

The overview matrix options offered in the user interface layer (Fig. 3) are a series of pre-programmed complex queries involving both biotic and abiotic datasets.

C. *Impact of informatics on microbial ecology*

The development of automated approaches to gathering molecular biology data has had a dramatic impact on the field of microbial ecology [12]. The bioinformatics tools that are necessary to analyze and interpret large datasets are having a similarly dramatic impact on the productivity of microbial ecology research. Automation of queries alone saves countless hours of researcher time. The approach described here allows us to leverage the utility of a costly DNA sequence dataset over several microbial community studies, conserving the resources (time and money) that would be necessary to complete a fine-scale phylogenetic analysis for each field study. These bioinformatics tools may also open new avenues of research as users mine the microbial community dataset for patterns that suggest the role of specific microbial populations in ecosystem processes. Sharing of microbial community datasets with researchers examining microbial populations in different environments may enhance inquiry related to microbially mediated ecosystem processes, thus strengthening the case for developing metadata standards among such datasets.

V. Conclusion

Information management and bioinformatics technologies have greatly aided ecological research efforts. The growing synergy between the disciplines of natural and computer sciences is changing the landscape of data collection, management and analysis. The discipline of microbial ecology lags behind traditional ecological disciplines in this regard. While microbial ecology researchers are comfortable with the technology necessary to gather and analyze their data, there are currently few informatics resources available to enable synthesis of these data with environmental conditions or ecosystem processes. The integration of molecular biology and environmental data described here can serve as a framework to improve analysis of existing data and may inspire future research integrating environmental microbiology datasets with large-scale studies.

References

[1] http://www.sun.com
[2] http://www.oracle.com
[3] http://lter.limnology.wisc.edu
[4] http://www.apache.org
[5] http://jakarta.apache.org/tomcat
[6] Smith, D. J., B.J. Benson, and D. F. Balsiger, " Designing web database applications for ecological research." *Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics. Vol VII: Informatics Systems Development II*, July 14-18, 2002
[7] http://java.sun.com
[8] http://java.sun.com/products/jsp
[9] http://ncbi.nlm.nih.gov/BLAST/
[10] http://rdp.cme.msu.edu/
[11] http://www.arb-home.de
[12] Venter, J.C., K. Remington, J.F. Heidelberg, et al. “Environmental genome shotgun sequencing of the Sargasso Sea.” *Science* 304:66-74.