Efforts to Link Ecological Metadata with Bacterial Gene Sequences at the Sapelo Island Microbial Observatory

Wade M. SHELDON
Department of Marine Sciences
University of Georgia
Athens, Georgia 30602, USA

and

Mary Ann MORAN
Department of Marine Sciences
University of Georgia
Athens, Georgia 30602, USA

and

James T. HOLLIBAUGH
Department of Marine Sciences
University of Georgia
Athens, Georgia 30602, USA

ABSTRACT

The existence of public databases for archiving genetic sequence data, such as GenBank and the Ribosomal Database Project, coupled with the availability of standardized sequence alignment and comparison tools has led to rapid advances in the field of bacterial genetics and systematics. Many microbial ecologists now routinely submit gene sequences obtained from environmental isolates, clones, and bands excised from electrophoretic gels to public sequence databases. As the amount of environmental sequence data in these systems has increased, ecologists have begun using sequence databases for broader classes of studies, such as biogeography and community ecology. Unfortunately, the general lack of documentation and data quality control standards has resulted in many sequences being entered without appropriate metadata, effectively orphaning records from their ecological context information and making comparisons impossible.

In order to address the shortcomings of public sequence databases, an independent 16S rRNA sequence database was recently developed at the Sapelo Island Microbial Observatory (SIMO) in Georgia, USA. The database was created to store complete information from all SIMO research activities using a hierarchical structure designed to reflect the actual flow of information from sample collection through final publication. By incorporating key fields from external databases, such as GenBank, the SIMO database is able to serve both as an independent research tool for SIMO scientists and as a reference source for SIMO data stored in other databases.

Keywords: Database, Metadata, 16S rRNA, Bacteria, Genomics, GenBank, Microbial Observatory

1. INTRODUCTION

The creation of public repositories for archiving genetic sequence data, such as GenBank and the Ribosomal Database Project, coupled with the availability of standardized sequence alignment and comparison tools has led to rapid advances in the field of bacterial genetics. In most research laboratories, comparison of 16S rRNA gene sequences has replaced traditional culture and phenotype-based methods for bacterial classification (Giovannoni and Rappe, 2000; Gonzalez et al., 2000). Most major microbiology journals, such as Microbiology and Applied and Environmental Microbiology, now require GenBank accession numbers as a pre-requisite for all sequences published in manuscripts. Not surprisingly, public

sequence databases have been growing at exponential rates in recent years. As of February 2002, there were approximately 15,465,000 sequence records in GenBank alone (NCBI, 2002).

Many microbial ecologists now routinely submit gene sequences obtained from environmental isolates and even bands excised directly from electrophoretic gels to public sequence databases for archival and analysis. As the amount of environmental sequence data has grown, these investigators have also begun to apply these databases to new classes of ecological studies, such as microbial community ecology and biogeography. Unfortunately, these efforts are often hindered by the lack of robust documentation standards for environmental sequence data. In the absence of standards, most sequences are entered without appropriate ecological and methodological metadata, effectively orphaning records from their research context information and making comparisons impossible.

The lack of complete metadata is particularly problematic for sequences obtained by direct amplification of DNA obtained from environmental samples (i.e. environmental sequences), because unlike sequences from bacterial isolates there is usually no reference material available for independent verification. In addition, recent reports of significant data quality control problems (Karp et al., 2001) and lapses in sequence format enforcement (Karp, 2001) for GenBank records further underscore the need for improving metadata standards for genetic sequences used for ecological research.

2. SIMO 16S rRNA DATABASE

A workshop was held in August 2000 during the LTER All-Scientists meeting to facilitate inter-site comparisons of microbial community composition and biogeography (Hollibaugh and Priscu, 2000). The workshop resulted in a draft proposal for the creation of an integrated environmental sequence database as an extension or companion to the existing public database resources to address the issues listed above. While this proposal has yet to be funded or implemented, many of the principles outlined were recently incorporated in a new 16S rRNA genetic sequence database developed at the Sapelo Island Microbial Observatory (SIMO) in Georgia, USA.

The SIMO database was designed to store complete information from all SIMO research activities in a hierarchical structure modeled after actual laboratory workflow patterns. Each successive step in the data hierarchy references the preceding steps, so that the full research context of all data is maintained from sample collection through analysis and final publication. Figure 1 illustrates the conceptual design of the database, and Table 1 lists the database table entities used to implement the model with commercial relational database management software (Microsoft SQL Server™). The full entity-relationship diagram is also available on the SIMO World Wide Web site (http://simo.marsci.uga.edu/public_db/).

The 'Samples' table represents the top of the database hierarchy, and also functions as the primary link between sequence data, spatial (i.e. geographic) and environmental information, and ancillary analyses and data. The subsidiary data tables, such as 'Source', 'Sequence', and 'SeqComparison', store both research data and laboratory management details, including storage amounts and locations, user-assigned codes or aliases, and processing notes. The inclusion of these laboratory attributes allows the database to serve an additional role as a laboratory information management system.

Metadata are primarily associated with individual data records by references to fields in related lookup tables (e.g. study sites, macro-environments, microenvironments, standard methodology, personnel records, storage locations). This design simplifies data entry (see below), minimizes redundancy, and encourages research standardization. In addition, the inclusion of metadata fields as foreign keys in data tables facilitates fine-grained queries and data sub-setting by metadata category, which would be far less effective for free-text fields without controlled vocabularies.

In addition to SIMO data, key fields from external databases are also stored in the database (e.g. GenBank accession numbers, Georgia Coastal Ecosystems LTER Project sampling site codes, bibliographic citations). These links provide access to additional supporting information and dramatically increase the scope of the SIMO database.

3. DATA ENTRY AND INTEGRITY

Both the database model and user interface are designed to maximize the quality of data and metadata stored in the system. Declarative referential integrity constraints are defined for all table relationships to prevent orphaned records and duplicate entries. Strict content control of key organizing terms is maintained by limiting data entry to lists of values that reference fields in lookup tables that are maintained exclusively by SIMO administrators and database managers. When list-based entries are not possible, input masks and validation functions are used to ensure that data entered by users is appropriate for the corresponding field.

A web-based data entry application allows users to enter data in a series of discrete steps (e.g. sample information, clone and isolate information, sequence data, phylogenetic information), which supports the discontinuous nature of molecular genetics research and allows delegation of data entry tasks to multiple students and technicians. The web application automatically generates appropriate hyperlinks to support both sequential data entry (e.g. samples through phylogenetic information) and multiple data entry into a single table (e.g. sequences from one or more clones) in any combination. The rigid application of referential integrity and hierarchical design of the database ensures that data cannot be added unless supporting information is already in place.

A separate administrative application was also developed using Microsoft™ Access 2000. This application allows authorized SIMO staff and database managers to update metadata and lookup tables, restore records deleted by mistake, and control the content of web application selection lists by setting values in management fields included in database tables. A granular, role-based security scheme is used to control access to all database entities, which prevents unauthorized personnel from using the administrative application to modify the database.

4. SIMO METADATA

Various metadata standards and data formats have been developed to facilitate exchange of genetic sequence data among researchers and various computer programs. Four of the most popular formats are as follows:

- FASTA file format[1]
- GenBank flat-file format [2]
- Biopolymer Markup Language (BIOML)[3]
- Bioinformatic Sequence Markup Language (BSML)[4]

These standards describe the formatting of the sequence data and provide varying levels of support for basic research metadata (e.g. researcher contact information, comments on methodology and analysis, and sequence annotation), but none provide the level of detail required to support ecological research and syntheses. In particular, support for specific research origin descriptors (e.g. site characteristics, sampling design, detailed methodology) and supplemental descriptors (quality control measures, reference materials, publication history) are totally lacking (see Michener, 2000, for a discussion of ecological metadata requirements).

In the absence of an established ecological metadata standard for documenting bacterial gene sequences, metadata information is currently displayed in summary format along with sequence details on the SIMO web site. Dynamic hyperlinks are created to provide access to the corresponding sampling site information, contact information for the responsible researcher, and ancillary data available in the Georgia Coastal Ecosystems LTER Project database.

Sequence data is also provided in the terse FASTA file format, with the SIMO unique identifier and web site URL listed in each sequence comment line. We are currently evaluating the suitability of emerging standards for ecological metadata, such as EML (Ecological Metadata Language; http://ecoinformatics.org/), for providing complete metadata for SIMO sequences in a more autonomous and computer-parseable format suitable for both data archival and dissemination.

5. FUTURE DIRECTIONS

The initial implementation and population of the SIMO database has been very successful, but a

---

[1] National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/BLAST/fasta.html)
[2] National Center for Biotechnology Information (http://www.ncbi.nih.gov/Genbank/)
[3] Proteometrics, LLC (http://www.bioml.com/BIOML/)
[4] LabBook, Inc (http://www.labbook.com/products/xmlbsml.asp)

number of potential barriers remain to realizing its full potential. We initially planned to provide SIMO unique identifiers and database URL pointers with all data submitted to GenBank to support bi-directional searching based on paired database keys. Unfortunately, our attempts to date have met with mixed success. NCBI GenBank personnel frequently remove SIMO pointer information from submissions made using BankIt (i.e. GenBank web form submission tool) based on poorly qualified criteria, such as concern over long-term URL stability. In contrast, sequences submitted as batches using the stand-alone Sequin program often retain their pointer information, indicating they receive less individual scrutiny.

To work around this limitation, efforts are now underway to list SIMO as a non-bibliographic sequence data provider in LinkOut, a referral system paired with GenBank as part of Entrez (http://www.ncbi.nlm.nih.gov/Entrez/). LinkOut is specifically designed to provide hyperlinks to ancillary information for sequences stored in GenBank, making it an ideal candidate for managing external SIMO data links. Unfortunately few microbial ecologists use the Entrez system because of its traditional association with medical journals (e.g. PubMed) rather than scientific research databases.

A combined approach that includes listing of SIMO sequences in LinkOut and a researcher education campaign will probably be required to accomplish a functional linkage between SIMO sequences stored in GenBank and the SIMO 16S rRNA database.

6. CONCLUSION

The vast amount of information available in public gene sequence databases represents an important resource for microbial ecologists studying biogeography, community structure, and biodiversity. Analytical tools and techniques that are currently available to mine these databases allow researchers to integrate data from across the world and ask truly large-scale questions that could never be addressed by a single researcher or laboratory. Without access to high quality ecological metadata, though, the scope of these studies will be severely limited.

The SIMO 16S rRNA database provides a working model of a system for coupling genetic information with ecological metadata. The query pages on the
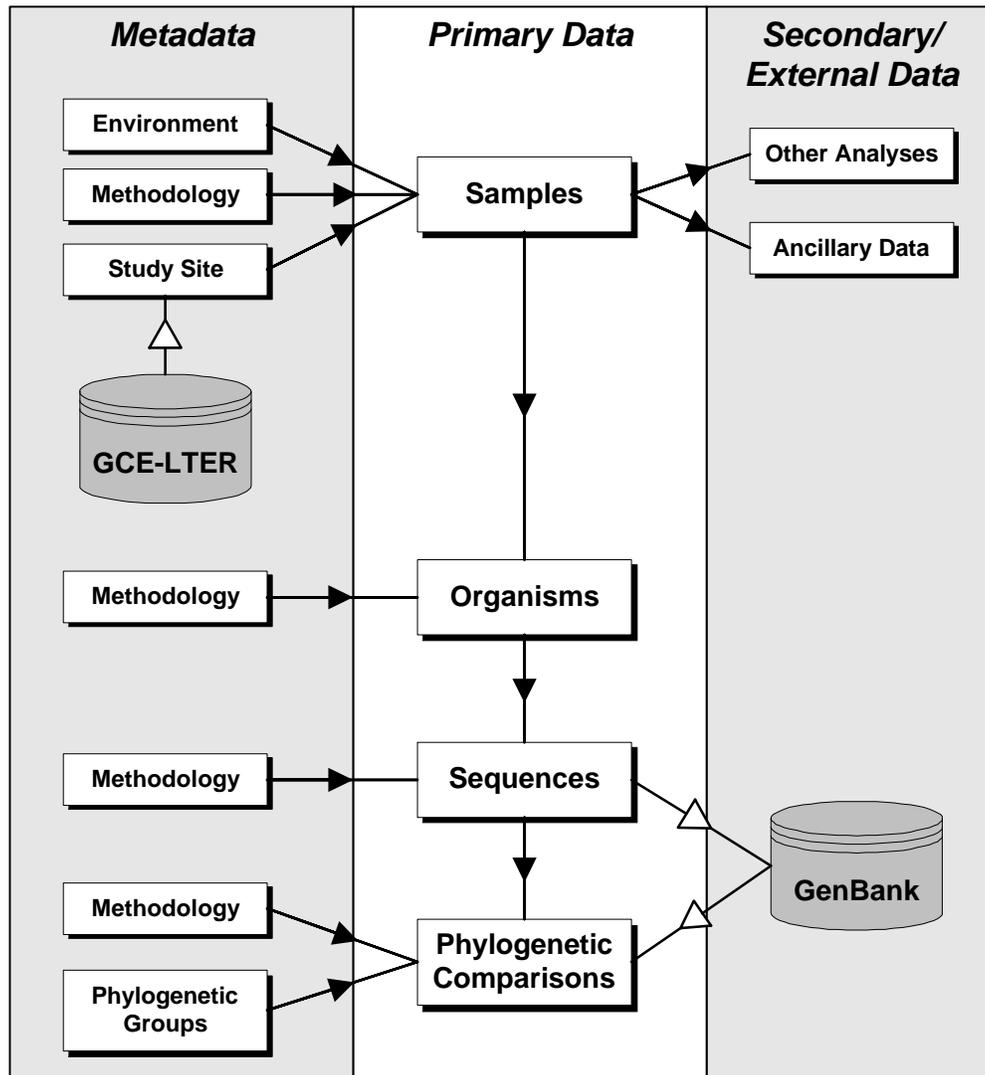
SIMO web site best illustrate the potential power of this model, providing researchers the ability to search for sequences by various phylogenetic characteristics, ecological characteristics, or an intersection matrix of combined characteristics (http://simo.marsci.uga.edu/public_db/). A national database incorporating key features of the SIMO database could dramatically enhance the potential for large-scale studies on bacterial biogeography and community structure.

7. ACKNOWLEDGEMENTS

8. LITERATURE CITED

Giovannoni, S. and Rappe, M., 2000. Evolution, Diversity, and Molecular Ecology of Marine Prokaryotes. In: D.L. Kirchman (Editor), Microbial Ecology of the Oceans. Wiley-Liss, New York, pp. 47-84.

Gonzalez, J.M., Simo, R., Massana, R., Covert, J.S., Casamayor, E.O., Pedros-Alio, C. and Moran, M.A., 2000. Bacterial community structure associated with a dimethylsulfoniopropionate-producing North Atlantic algal bloom. Applied and Environmental Microbiology, 66(10): 4237-4246.

Hollibaugh, J.T. and Priscu, J.S., 2000. Microbial Biogeography: Cross-site comparison of aquatic systems, LTER All-Scientists Meeting, Snowbird, Utah, USA.

Karp, P.D., 2001. Many GenBank entries for complete microbial genomes violate the GenBank standard. Comparative and Functional Genomics, 2(1): 25-27.

Karp, P.D., Paley, S. and Zhu, J., 2001. Database verificiation studies of SWISS-PROT and GenBank. Bioinformatics, 17(6): 526-532.

Michener, W.K., 2000. Metadata. In: W.K. Michener and J.W. Brunt (Editors), Ecological Data - Design, Management and Processing. Methods in Ecology. Blackwell Science Ltd., London, pp. 92-116.

NCBI, 2002. GenBank Overview. National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/Genbank/Genbank Overview.html).

figure 1. Conceptual diagram of the Sapelo Island Microbial Observatory 16S rRNA database illustrating the relationships and directions of information flow between metadata, primary data, and secondary data entities (note that some related metadata tables are combined for clarity). Solid arrows represent internal relationships with referential integrity constraints, hollow arrows represent external relationships based on shared keys, and cylinders represent external databases. GenBank is the NIH genetic sequence database, and GCE-LTER is the Georgia Coastal Ecosystems LTER Project metadata database.

Table 1. Relationships among the primary tables comprising the SIMO 16S rRNA sequence database. The full entity-relationship diagram is available on the WWW at http://simo.marsci.uga.edu/public_db/.

Data Tables	Descriptions	Related Tables	Related Table Descriptions
Samples	Information about physical samples collected for DNA extraction or bacterial isolation	Site	Sampling site details, including geography
		Environment	Macro-environment designation and description
		MicroEnv	Microenvironment designation and description
		Zone	Marsh zone designation and description
		Sampling	Sampling methodology description
		Storage	Storage method and location details
		People	Contact information for the responsible party
Source	Information about DNA sources (bacterial isolates and vectors into which extracted and PCR-amplified DNA was cloned)	**Samples**	Sample from which the sequence source was derived
		SourceType	Category of sequence source (lookup table)
		SourceTechnique	Methodology used for isolation or cloning and maintaining the organism
		People	Contact information for the responsible party
		Storage	Storage method and location details
Sequence	Information about nucleotide sequence data obtained by DNA analysis of 16S rRNA genes	**Source**	Source from which the DNA sequence was obtained
		SeqAnalysis	Sequence analysis methodology and instrumentation
		Primers	Description and vendor information for DNA primers used in the analysis
		People	Contact information for the responsible party
SeqComparison	Information about phylogenetic comparisons between SIMO sequences and other sequences published in GenBank, made using sequence alignment and analysis programs such as BLASTN and FASTA	**Sequence**	Sequence used for comparisons
		Programs	Information about the program used for the comparison
		BactDivision	Bacterial division (taxonomic lookup table)
		People	Contact information for the responsible party
Phylogeny	Information about the most closely related organisms identified by sequence comparisons	**SeqComparison**	Comparison reference