J. Microbiol. Biotechnol. (2006), **16**(2), 178–183

JOURNAL OF MICROBIOLOGY AND BIOTECHNOLOGY

Development of a Meta-Information System for Microbial Resources

YU, JAEWOO[1,2], WONHYONG CHUNG[1], TAE-KWON SOHN[1], YONG-HA PARK[1,2], AND HONGIK KIM[1]*

[1]*proBionic Corp., Daejeon 305-333, Korea*
[2]*Biological Resource Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 305-333, Korea*

Received: January 29, 2005
Accepted: October

Abstract Microbes are one of the most important bioresources in bioindustry and provide high economic values. Although there are currently about 6,000 bacterial species with validly published names, microbiologists generally assume that the number may account for less than 1% of the bacterial species present on Earth. To discover the remaining species, studies of metagenomes, metabolomes, and proteomes related to microbes have recently been carried out in various fields. We have constructed an information system that integrates various data on microbial resources and manages bioinformation to support efficient research of microorganisms. We have designated this system "Bio-Meta Information System (Bio-MIS)." Bio-MIS consists of an integrated microbial resource database, a microbial resource input system, an integrated microbial resource search engine, a microbial resource online distribution system, a portal service, and management via the Internet. In the future, this system is expected to be connected with various public databases. We plan to implement useful bioinformatics software for analyzing microbial genome resources. The Web site is accessible at http://biomis.probionic.com.

Key words: Microbial resource, information system, metagenome, metabolome, proteome

Great strides have recently been made in identifying and characterizing the staggering diversity of microorganisms that conduct primary and secondary production, nutrient transformation, and mineralization processes that underlie ecosystem and regional biogeochemical, trophodynamic, and ecological change [16, 21]. At the molecular level, however this diversity is poorly represented. Estimates suggest that less than a small fraction of prokaryotes have been isolated, and representatives of only about 10 to 15% of described species are held in service culture collections [3]. It is important that we understand the whole cell network of microbes in order to discover useful biomolecules and new functional bioprocesses from hither-to cultivated and unculturable microorganisms [3, 20]. The growth of biological information has accelerated by the development of high-throughput analyses such as DNA microarray. This technology provides clues to the discovery of novel genes, and biomolecules and bioprocesses that have not yet been revealed owing to the lack of proper information [5, 15]. With this explosive increase of biological information, it becomes necessary to store and efficiently link it with other related information [2].

Recent years have seen an explosion in the amount of available biological data. More and more genomes are being sequenced and annotated, and protein and gene interaction data have accumulated [18]. In the case of bacteria, 160 complete genomes have been sequenced and 2 are in progress as of June 2004 [17, 22]. In addition, about 50,000 16S rDNA sequences are available in the ribosomal database [4, 23]. This increasing flow of data has thus far been managed through the creation of numerous biomedical databases coupled to search engines, query interfaces, and analysis tools [8]. Databases of biological knowledge have grown from a cottage industry that was only of interest to a few specialized disciplines, to become essential resources that are used daily by biologists around the world [18]. In the area of bioinformatics, search strategies are based upon data collection and storage and the mining of databases in order to generate knowledge [19]. Although there are more than 200 microbiology-related databases, it is difficult, if not impossible, to find answers to questions that rely on the use of integrated information from even a few databases. Hence, data integration is a desideratum [3]. A prototype integrated microbial database (IMD) project was launched by the Center for Microbial Ecology at Michigan State University in 1997 [13]. Among the various types of information, the management of sequence information has been studied widely. Sequences constitute the basic data obtained from analyzing an organism and are an effective means to normalize, compare, and formalize the

*Corresponding author
Phone: 82-42-862-1320; Fax: 82-42-862-1315;
E-mail: hikim@probionic.com

expression of analyzed data [10]. However, the biological information related to microbial resources, such as metagenome, metabolomes, and proteomes, is difficult to normalize, compare, and formalize, because the expression of resources may be different depending on the observer.

"Bio-Meta Information System (Bio-MIS)" was developed as a means to integrate all information retrieved from a biocomplex system and to share heterogeneous microbial data from an organized microbial ecosystem. In this manner, it becomes possible to understand the microbialcommunity and cell mechanism. A flexible and robust data structure is integral for successful management of such a complex information system. The query interface provides easy access to this information for the research community. In this paper, we demonstrate an efficient information management system for various microbial data.

MATERIALS AND METHODS

System Environments

The hardware environments for the development of Bio-MIS were an HP Prolient ML370 computer for the server, and a Pentium IV IBM PC for data reprocessing, data extraction, and development of programs. Software environments for development were Microsoft Windows 2000 Advanced Server for the operating system, IIS (Internet Information Server) 5.0 for the Web-server system, and MS-SQL 9.0 for the Database Management system. Bio-MIS was written in Visual Basic Script and HTML (hyper text markup language). We used HTML to make the Web interface of the information input system and portal site Web pages. ASP (Active Server Page) technology was used for the database access module and data retrieving program.

The Data Structure

Bio-MIS has been adopted as a relational database model. Biological information organized in a table form may have some inefficient data if it is stored in the database as a table directly. In order to maximize the flexibility of the data structure and minimize duplicated data, we separated the biological information in terms of data form, data content, data access information, and auxiliary information. Data form corresponds to field information if the information is made up of a single table. Data content is a value that is stored in the table. Data access information is information about the owner, access priority, etc. Finally, auxiliary information for a data form is used when the form needs additional data such as a list type. The main parts of the data structure are shown in Fig. 1, where each part is represented as follows.

"'Data form" has the information of a data type and the information for constructing a table, which consists of seven fields. "Idx" is the index number of the data form,

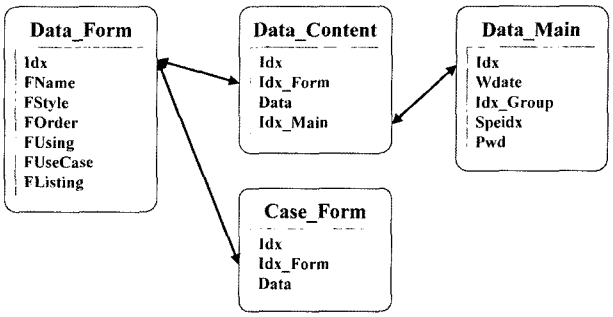

Fig. 1. Data structure of the Bio-MIS database.

and is assigned automatically. "FName" is the name of the data type, and is used to display the field name of a table. "FStyle" represents the type of data, and decides how the interface layer interprets data_content. "FOrder" indicates the field's position when the table is constructed. "FUsing" checks whether the data form is used in showing the table, and enables the creation of reserved data forms. "FUseCase" indicates that the case list related to its FStyle is used in representing the field. Finally, "FListing" determines whether the data form is used for making a summary of the data, or showing a list of the data. "Data Content" contains the information of a value of a data form, and consists of four fields. It represents the value of a record and field pair of a table. "Idx" is the index number of the data content. "Idx_Form" shows the data form, and is associated with the content. "Data" field represents the content of data. "Idx_Main" is the link to the associated table information. "Data main" contains information about the table itself, the owner, access priority, etc., and consists of five fields. "Idx" is the index number of the data main, and is assigned automatically. "Wdate" shows the last update date of the table. "Idx_group" is the index of the table, which belongs to a predefined group. "Speidx" shows the index of the user, who has priority of write and modify. "Pwd" denotes the password for accessing a table. "Case_Form" has auxiliary information of the list style of a data form. In the case of a data form concerning air requirement, additional data such as "aerobic" and "anaerobic" are needed. "Idx" is the index number of data content. "Idx_Form" shows the data form, which is associated with the content. "Data" contains additional data as referred to in "Case_Form."

RESULTS

Integrated Database System

Biological information consists of heterogeneous data types. For example, metagenome information consists of clone name, type of clone, vector used, size of the metagenome, sequence, etc. The Bio-MIS database system decomposes such information into meta-information (data type) and

content (data value), and stores them separately. The interface of the database provides users with access to completely composed information from meta-information and contents. Bio-MIS can store many kinds of information and decides relationships between data fields, which enables information integration and efficient integrated searching.

Integrated Search Engine

The search engine provides the function of searching for the data of interest through relationships between databases and the contents of the databases. Retrieving data from other related databases aids in obtaining information that falls outside of well-known areas. For example, if users wants to know whether metagenome resources are related to a known protein, they can find the information from the search results for that protein.

Bio-Meta Information System

In the practical studies involving biological resources such as integrating biological information and related resources, information pertaining to a microbe such as the strain name, isolation data, properties, and its isolated clone need to be handled at the same time. Furthermore, the type and sort of information for describing biological resources can change during the time it is being used.

Therefore, in order to cope with these problems, an integrated system for managing biological information and distributing biological resources with a dynamic formatted database is required. Based on these considerations, we designed an information system that offers the following features: 1) efficient inputting of biological information; 2) integrated information search from various biological data; 3) online distribution of bioresources; and 4) capacity to change the schema of the database in use. We have named our system "Bio-Meta Information System (Bio-MIS)", because it decomposes the biological information into meta-information and contents. The system consists of an "Integrated database system," a "Biological information input system," an "Integrated search engine," a "Biological resources on-line distribution system," and a "Bio-MIS portal service." We describe these five components of the Bio-Meta Information System in the following section. Figure 2 shows the system organization of Bio-MIS for microbial resources.

Biological Information Input System

Bio-MIS offers a Web-based input system for biological information, which is stored in the meta-database. The input system consists of a Web-based input form, input form interpretation, and input form management. The first component is the Web-based input form, which has the role of communicating with the user about adding, removing, and modifying information via the Bio-MIS portal service. The second component is the input form interpretation, which interprets user inputs received from the Web as data types of the database. Because our database's data value

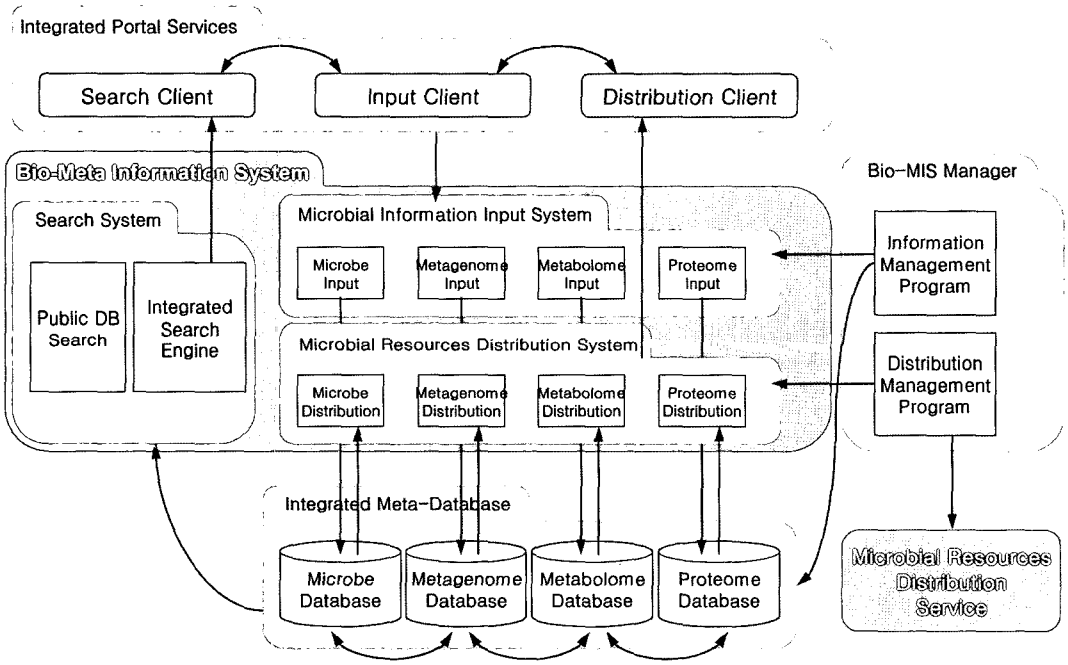

Fig. 2. A schematic representation of Bio-MIS for microbial resources management, consisting of microbe, metagenome, metabolome, and proteome databases.
The Bio-MIS framework can also be implemented for various biological data sets.

contains no type information, the value should be stored or retrieved by referring to its meta-information. Hence, the input form interpretation has an important role of linking the user and the flexible database in Bio-MIS. The last component is the input form management, which defines the format of the input form and the summary of a data record. An identification number is used as the title of the record when a database is displayed as a table.

Biological Resources Online Distribution System

Bio-MIS adopts the concept of an online shopping mall. Online shopping malls display goods to customers through Web pages, or provide a search function to customers who want to find something. Customers survey the information and subsequently make buying choices. The goods are transferred to customers through an automated procedure, and customers can monitor the status of the shipment. Our online distribution system has two functions; one that distributes biological resources to customers via an online shopping mall, and another that informs a distribution manager of a customer request and how to respond to that request. Additionally, the "Resources management program" helps the resources manager manage the total stock of resources, the state of newly deposited resources, etc.

Bio-MIS Portal Service

The Bio-MIS portal service provides an integrated Web-based user interface for the information input system, search engine, and distribution system. It also provides news, an online community service, a messaging function between users, as well as other functions. Because users of Bio-MIS have common interests (e.g., users of Bio-MIS for microbial resources may have a shared interest in news about newly discovered metabolism), the Bio-MIS portal service provides information about biological resources to users.

Database for Microbial Resources

Bio-MIS provides an integrated service for gathering and distributing microbial resources to various science fields such as microbial diversity, metagenomes, metabolic engineering, and genomic applications. We have applied our pre-designed system, "Bio-MIS," for microbial resources collected from a microbe resource bank, "Microbank" in the Microbial Genomics and Applications Center in South Korea. Figure 3

Name	Display type	Order	Usage	For List
Microbe name	Edit	1	True	True
Scientific name	Edit	2	True	False
Source	Edit	3	True	True
Place	Edit	4	True	False
Date	Edit	5	True	False
Culture Temperature	Edit	6	True	False
Culture PH	Edit	7	True	False
Culture Medium	Edit	8	True	False
Custom Culture Medium	Text	9	True	False
Aerobic	Combo	10	True	False
Characteristics	Combo	11	True	False
Other Characteristics	Text	12	True	False
16S rDNA sequence	Text	13	True	False
Owner	Edit	14	True	True

(A) Microbe information database

Name	Display type	Order	Usage	For List
Clone name	Edit	1	True	True
Clone type	Combo	2	True	False
Used vector	Combo	3	True	False
DNA sample	Edit	4	True	True
Extraction Place	Edit	5	True	False
Extraction Date	Edit	6	True	False
Metagenome Size	Edit	7	True	False
Metagenome Charateristics	Text	8	True	False
Metagenome sequence	Text	9	True	False
Owner	Edit	10	True	True

(B) Metagenome information database

Name	Display type	Order	Usage	For List
ML number	Edit	1	True	True
Plate number	Edit	2	True	True
Addr. on 96well	Edit	3	True	True
Strain	Edit	4	True	True
Producing strain	Edit	5	True	True
Volume	Edit	6	True	True
Activity	Edit	7	True	False
Solvent	Edit	8	True	False
Structure	Edit	9	True	False

(C) Metabolome information database

Name	Display type	Order	Usage	For List
Microbe name	Edit	1	True	True
LOCUS	Edit	2	True	True
Common name	Edit	3	True	True
TIGR ID	Edit	4	True	True
GBID	Edit	5	True	True
Protein sequence	Text	6	True	False
Gene sequence	Text	7	True	False
Activity 1	Edit	8	True	False
Activity 2	Edit	9	True	False
Activity 3	Edit	10	True	False

(D) Proteome information database

Fig. 3. The structure of databases for microbial resources.
The database consisted of four subdatabases. The five data fields are the major components of the "Data_Form" table, which forms a table of a genome resource.

shows the implemented fields of each resource. This system began operation in October 2003, and provides information of collected microbial resources and distributes these resources to interested scientists. The Web site of Microbank is http://www.microbank.re.kr.

Bio-MIS for microbial resources currently has 31,178 records of strain data, 76 records of metagenome data, 880 records of metabolome data, and 361 records of proteome data. First, the Metagenome section provides 12 fields of data, including identification number, clone name, clone class, vector name, metagenome size, metagenome features, and provider name. Second, the Metabolome section provides 11 fields of data, including identification number, ML number, strain number, sample volume, activity description, solvent name, and structure. Third, the Proteome section provides 12 fields of data, including identification number, common name, strain name, locus, TIGR ID, GenBank ID, protein sequence, gene sequence, and activity description. Finally, the Strain section provides 16 fields of data, including strain name, species name, isolation source, isolation place, isolation date, culture temperature, proper culture pH, media composition, and 16S RNA sequence.

Discussion

The most popular database model is the relational model, which has gained favor over any other database model in recent years and has been implemented by many vendors commercially and noncommercially. Accordingly, we developed the database using MS-SQL, which is a database management system based on the relational model. In the following, we briefly describe the relational database. A relational database stores all data inside tables. All operations on data are conducted on the tables themselves or other tables are produced as a result. The rows from a relational table are analogous to a record, and the columns to a field. Each row is a set of columns with only one value for each. All rows from the same table have the same set of columns [12].

This is the best way to construct only one table for biological information, because a single table has no relation in the database. This means that there is no need for complex operations, and it is easy to intuitively find information from the table. It is difficult to define a table for biological information, particularly for ongoing research. Because biological information obtained from experiments may have various forms and value according to the measuring instruments or the development of new experiment methods.

These types of problems were encountered when we applied Bio-MIS to microbial resources. After the design of the database for microbes, metagenomes, metabolomes, and proteomes was completed, there was demand for modifying the type and the name of some data fields. Designing the data structure of Bio-MIS as one table for intuitively retrieving information would have been extremely time-consuming and difficult. During the period of operating Microbank over several months, more accurate information of the culture conditions in the microbe's database, such as "Optimal Culture Temperature" for thermophiles, was found to be necessary. Our system's flexibility successfully adapted to this case. The manager of Microbank simply added that field to the database via the management program of Bio-MIS.

Although the database of Bio-MIS provides a flexible data structure that can be easily modified and managed, the meta-information should be carefully defined when constructing a new database in Bio-MIS. The content of information is not entered into the existing records when a field is added in the constructed database, and the newly defined field's information is stored in the Data_Form. Hence, the contents of the type should manually be filled in the Data_Content. For example, when we added the "Optimal Culture Temperature" field in the microbe database, it was not possible to automatically add the field's content into the existing records, because the field does not have a default value. We decided to display a "Null" value when searching this field and subsequently informed the user who registered the data that it had been updated. This solution did not cause any problems, because the newly added field is relatively less important in the database.

This problem can be resolved by an effective initial definition of meta-information and the operating policy of the system rather than by an upgrade of the data structure or system. Although the Bio-MIS cannot solve this problem, it accelerates the performance of the resource management system by carefully defining core meta-information and simply, defining auxilary meta-information because a change of core meta-information may cause a problem whereas a change of auxilary meta-information rarely causes any problem.

In summary, we have shown that the data structure of Bio-MIS has the flexibility of changing fields of information after the database has been constructed. The Bio-MIS was successfully applied to a microbial resource management system and has served as a framework of Microbank's services. The current version of data structure has a simple linkage between related data, and this linkage should be defined maunally. We are currently involved in automatically constructing relationships between data fields of each item of information, which enables information integration and efficient integrated search. Finally, we hope to build a database framework for biological data integration.

Acknowledgments

This work was supported by the 21C Frontier Microbial Genomics and Application Center Program, Ministry of

Science & Technology (Grant MG02-0501-001-1-0-0), Republic of Korea. We would like to thank Drs. Jung-Hoon Yoon, Choong Hwan Lee, and Hyeon-Su Ro at the Microbial Genomics & Application Center for their support and suggestions.

REFERENCES

1. Ahn, G.-T., J.-H. Kim, K.-M. Kang, M.-J. Lee, and I.-S. Han. 2004. BioPlace: A web-based collaborative environment for effective genome research. *J. Microbiol. Biotechnol.* **14:** 1081–1085.
2. Bruskiewich, R. M., A. B. Cosico, W. Eusebio, A. M. Portugal, L. M. Ramos, M. T. Reyes, M. A. B. Sallan, V. J. M. Ulat, X. Wang, K. L. McNally, R. S. Hamilton, and C. G. McLaren. 2003. Linking genotype to phenotype: The international rice information system (IRIS). *Bioinformatics* **19:** i63–i65.
3. Bull, A. T., A. C. Ward, and M. Goodfellow. 2000. Search and discovery strategies for biotechnology: The paradigm shift. *Microbiol. Molec. Biol. Rev.* **64:** 573–606.
4. Cole, J. R., B. Chai, T. L. Marsh, R. J. Farris, Q. Wang, S. A. Kulam, S. Chandra, D. M. McGarrell, T. M. Schmidt, G. M. Garrity, and J. M. Tiedje. 2001. The ribosomal database project (RDP-II): Previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res.* **29:** 123–125.
5. Demain, A. L. 2001. Genetics and microbiology of industrial microorganisms. *J. Ind. Microbiol. Biotechnol.* **27:** 352–356.
6. Desvaux, M. 2004. Mapping of carbon flow distribution in the central metabolic pathways of *Clostridium cellulolyticum*: Direct comparison of bacterial metabolism with a soluble versus an insoluble carbon source. *J. Microbiol. Biotechnol.* **14:** 1200–1210.
7. Do, J. H., M. J. Anderson, D. W. Denning, and E. Bornberg-Bauer. 2004. Inference of *Aspergillus fumigatus* pathways by computational genome analysis: Tricarboxylic acid cycle (TCA) and glyoxylate shunt. *J. Microbiol. Biotechnol.* **14:** 74–80.
8. Gelbart, W. M. 1998. Databases in genomic research. *Science* **282:** 659–661.
9. Hong, S. H., S. Y. Moon, and S. Y. Lee. 2003. Prediction of maximum yields of metabolites and optimal pathways for their production by metabolic flux analysis. *J. Microbiol. Biotechnol.* **13:** 571–577.
10. Janssen, P., A. J. Enright, B. Audit, I. Cases, L. Goldovsky, N. Harte, V. Kunin, and C. A. Ouzounis. 2003. COmplete GENome tracking (COGENT): A flexible data environment for computational genomics. *Bioinformatics* **19:** 1451–1452.
11. Kim, T.-W., S.-H. Jung, J.-Y. Lee, S.-K. Choi, S.-H. Park, J.-S. Jo, and H.-Y. Kim. 2003. Identification of lactic acid bacteria in kimchi using SDS-PAGE profiles of whole cell proteins. *J. Microbiol. Biotechnol.* **13:** 119–124.
12. Korth, H. F. and A. Silberschatz. 1991. *Database System Concepts*, pp. 53–91. 2nd Ed. McGraw-Hill Inc., 2 Penn Plaza, New York, U.S.A.
13. Larsen, N., R. Overbeek, S. Pramanik, T. M. Schmidt, E. E. Selkov, O. Strunk, J. M. Tiedje, and J. W. Urbance. 1997. Towards microbial data integration. *J. Ind. Microbiol. Biotechnol.* **18:** 68–72.
14. Nelson, K. E. 2003. The future of microbial genomics. *Environ. Microbiol.* **5:** 1223–1225.
15. Lee, J.-Y. and N. G. Lee. 2004. Transcriptional responses of human respiratory epithelial cells to nontypeable *Haemophilus influenzae* infection analyzed by high density cDNA microarrays. *J. Microbiol. Biotechnol.* **14:** 836–843.
16. Paerl, H. W. and T. F. Steppe. 2003. Scaling up: The next challenge in environmental microbiology. *Environ. Microbiol.* **5:** 1025–1038.
17. Peterson, J. D., L. A. Umayam, T. Dickinson, E. K. Hickey, and O. White. 2003. The comprehensive microbial resource. *Nucleic Acids Res.* **31:** 442–443.
18. Stein, L. D. 2003. Integrating biological databases. *Nature Rev. Genet.* **4:** 337–345.
19. Tiedje, J. M. 1995. Approaches to the comprehensive evaluation of prokaryotic diversity of a habitat, pp. 73–87. *In* D. Allsopp, R. R. Colwell, and D. L. Hawksworth (eds.), *Microbial Diversity and Ecosystem Function*. CAB International, Wallingford, U.K.
20. Torsvik, V. and L. Ovreas. 2002. Microbial diversity and function in soil: From genes to ecosystems. *Curr. Opin. Microbiol.* **5:** 240–245.
21. Yagisawa, M. 2000. Trends in exploratory research of microbial metabolite part 1; 50 years from the discovery of penicillin. *Bioscience and Industry* **58:** 12–17.
22. http://www.tigr.org/tigr-scripts/CMR2/CMR_Content.spl (The Comprehensive Microbial Resource).
23. http://rdp.cme.msu.edu/html/ (The Ribosomal Database Project).