BACKGROUND

The *E. coli* Genetic Stock Center (CGSC) database was designed in 1989 and implemented in 1990 with the purpose of serving the needs of the center and its present and future staff to use electronic records to search for genotypic combinations; to track maintenance and request records; to maintain information on pedigrees, allele numbers, gene descriptions, and map data, all tied to literature references; and to search for strains or other objects on the basis of combinations of any property of any of these objects. The collection currently includes about 7,000 strains, with combinations of mutant alleles ranging from 0 to 25 per strain, including subcollections of Hfr and Hfr Tn10 strains useful for mapping and a plasmid library of cloned genes from *Escherichia coli K-12*. The datasets which served to define the requirements and modeling efforts and to populate the initial segment of the database are the notebooks, card catalogs, and other records of Barbara J. Bachmann, curator of the Stock Center since 1973 (1). Examples of queries which guided the design of this database are (i) find a strain with an amber mutation in *lacZ*, which is F−, carries no amber suppressor, and is streptomycin resistant; (ii) show all Tn insertions and points of origin in the 4- to 10-min region of the map; (iii) find a strain with an *ompT* mutation or a deletion that includes *ompT*; (iv) show the extent of all the deletions or F-prime plasmid inserts in a given region; and (v) find pairs of strains that are isogenic except for a *polA* mutation.

Once the requirements analysis, conceptual and data modeling, and data management system choices reached a functional level, with thanks to guidance from a database advisory group (this group included database specialists R. Robbins, National Science Foundation [now at F. Hutchinson Cancer Research Center], J. Ostell, National Center for Biotechnology Information, National Institutes of Health, and T. Marr, Cold Spring Harbor Laboratory, and geneticists K. B. Low, Yale University, and K. Sanderson, University of Alberta), the implementation was turned over to Stanley Letovsky, Letovsky Associates (now at Johns Hopkins University), who collaborated in the refinement and enhancement of the datamodel and then fully implemented the database. The choice we had made was to implement a robust relational database as a backend, with a powerful query interface as frontend for CGSC use and flexible capabilities for future on-line access and frontends. In response to expressions of interest by many researchers about direct access to the information describing strains, mutations, genes, gene products, and map locations, particularly by persons wishing to examine strain genotypes, we have provided access to that information by several routes, implemented again by Stanley Letovsky and effected in 1992. Access is available on virtually any hardware setup; an x-windows interface will be more pleasant to use than a vt100 terminal, but both APT (Sybase toolkit) forms and gopher accommodate this lowest-common-denominator vt100 access. In addition, the Sybase database was quickly adapted to a queriable World Wide Web frontend by using Letovsky’s Genera tool for using object-oriented specifications to automatically generate Web frontends while either generating or utilizing a preexisting Sybase backend (http://cgsc.biology.yale.edu/genera.html; see below and reference 5). This style of access was made available in early 1994.

Since electronic “Help” and “How To ...” files can be maintained and frequently updated as part of the database maintenance effort, I will not set down a dated version of such information here. Rather, this chapter provides information on how to access the database and highlights some of the content and features, leaving the reader to explore further. Since the Request-Handling and Strain Maintenance segments of the database are not of public interest, the “guest,” read-only version of the database which is offered to the public does not include access to these tables and forms, and they are not discussed here; nor are the map utilities associated with the database presented here. Details about these and other aspects of
the database can be found elsewhere (2–5). Tables and diagrams of the current map are presented on the Web. There are three modes of access: gopher (Fig. 1), Sybase APT forms (Fig. 2 and 3), and Web/Genera (Fig. 4 and 5). All can be reached from the Web site http://cgsc.biology.yale.edu.

**FIGURE 1** Opening menu of the gopher server, followed by selection and entry of a Strains query (a strain containing *carB* in its description) and one of the records retrieved.

**MODES OF ON-LINE ACCESS**

**Gopher**

Gopher is a standard and simple query-matching protocol that can be accessed from any type of computer. Persons who routinely access the main gopher at the University of Minnesota or any other gopher providing a geographic or subject index will find the CGSC database under Connecticut or Biology subject lists. Anyone who does not routinely access gopher servers can telnet and use the guest log-in described below to reach either the gopher or forms interface. The opening menu (Fig. 1) provides some background information about using the CGSC database in the “1. About the CGSC Gopher” file, and, once the “2. CGSC Files and Database/” option is selected, further instruction is available in the “About the CGSC...” file.
Some general information about using wais-indexed gophers for querying is found in the file entitled “IUWais Search Features.” The choices for querying the gopher are Strains, Sites (genes, operons, chromosomal fragments, etc.), and Mutations, and a map list ordering genes by coordinate can also be examined. When a category is selected, a query box invites the user to enter a character string (* may be used as a wild card, and Boolean ORs and ANDs can be used as described in the “IUWais” file) and all records in that category containing that string will be returned. This is easy, quick, and satisfactory in many cases. Sometimes, however, the nature of the query or the retrievals will suggest that a more specific or complex query would give a more helpful return. In those cases, the APT forms or Mosaic frontends will be more appropriate.

**APT Forms**

The most sophisticated querying can be done with the APT forms interface, as described in the “How to Access” file. This can be reached by requesting a password from berlyn@cgsc.biology.yale.edu and logging in as “guest.” The Web interface also allows specific and fairly complex querying and has hypertext connections to other databases. It will be the third mode described.

Figure 2a shows a query for a pair of strains that are constitutive for lacI and isogenic except for a mutation in lacZ. The query returns a list showing that there are two lacZ mutant strains that meet those requirements (Fig. 2b); each strain can be examined by using the “Browse!” option on the menu bar above the list. When one or both of those strains are then selected by indicating “+” and exiting (X) the Select form, the strain description of a chosen strain is presented (Fig. 2c) and the “Isos!” command on the menu bar above the strain form will retrieve the isogenic parent or sibling that matches the strain (Fig. 2d). The “Report” option presents a more formalized version of the descriptions (Fig. 2e).

Figure 3 shows a query from the mutation form for all Tn’s and points of origin (PO’s) inserted between 9 and 13 min on the E. coli chromosome. First, a list of the qualifying mutations is returned. The Draw! option presents the information in graphic form. Had the user wanted to examine strains containing one or more of those Tn’s or PO’s, he or she would have started the query on the Strain form, placed the cursor in the Mutation field, and used the Sub! option to go to the mutation form and enter the query as shown. In that case, when the mutations of interest are selected, they are returned to the strain form and the strain query is executed, resulting in a strain or list of strains carrying the selected mutations (not shown).

**World Wide Web**

From any World Wide Web (also referred to as WWW or Web) frontend, such as NCSA Mosaic or Netscape, the CGSC database is accessed by selecting Open and providing the URL http://cgsc.biology.yale.edu/top.html. As with the APT frontend, a powerful query forms interface is provided for CGSC in the WWW version. The forms again allow the user to fill in the blank in the desired category, and the Help button explains how to specify “A or B” and “A and B” types of queries. Since the Strain form on the Web has fields very similar to those shown in Fig. 2, we will illustrate the Mosaic frontend by using queries on genes. Figure 4 illustrates the Genera/WWW-form queries simply specifying (i) all genes and operons between min 1 and 3 that are transcribed in a counterclockwise direction and (ii) genes that specify the enzyme isocitrate dehydrogenase (or similarly named enzyme, since the wild card “%” is used). A record for one of the genes returned from the latter query and the many hypertext connections associated with that record are shown in Fig. 5. The WWW frontend bases much of its charm and utility on the ability to establish hypertext links with any other database on the Web. Examples of calls to other databases, as
FIGURE 2 Querying through the APT forms interface for pairs of strains that are constitutive at a given locus (*lacI*) and isogenic except for mutation at a second locus (*lacZ*). Thin arrows point to entries in fields which specify the query, and broad arrows point to menu commands that execute the action resulting in the retrieval shown on the form immediately below. Panels: (a) query; (b) selection form, showing the name and identification numbers of strains satisfying the query; (c) strain description of strain selected; (d) strain description of the isogenic line retrieved via the "Isos!" option; (e) the two strain descriptions reformatted via the "Report" function.
well as other records within the CGSC database, are shown in Fig. 5, illustrating hypertext moves from the CGSC icd gene description to mutations and strains within the database and to external databases GenBank, SwissProt, MedLine, XLocus, MaizeDB, SaccDB, GDB, and FlyBase. Clicking on an underlined portion of the record will establish a hypertext link to a record for that object. Clicking on the name of the database rather than the record identifier will take the user to the home page for that database, which provides reference information about the database and its developers.

It is hoped that in the very near future, these links will provide communication between the E. coli and Salmonella (1; chapter 109, this volume) and other bacterial databases and that we will be moving back and forth through bacterial information in the same way that we now do between E. coli, maize, and other species.
NAME REGISTRY AND CROSS-SPECIES LINKS

On-line databases also provide an opportunity for a public rather than one-on-one personal access to a registry of gene symbols and allele numbers. The opening page of the WWW server presents a pair of query boxes for entering a gene symbol to see if it is a symbol currently or previously used and to examine the records of the gene information. (A similar result would be obtained by entering the symbol or part of the symbol into the Mutation field of the STRAIN form of the APT forms interface, where any current symbol will be retrieved and any synonym will be replaced with the current symbol and be found in the Synonym field of the Site and Mutation records.) A request to register a new symbol could be transmitted electronically or directed to the CGSC by phone or mail.

To facilitate hypertext links between similar loci in different organisms, we have developed a Generagenerated XLocus database (M. Berlyn and S. Letovsky, unpublished data). This has proven useful for sets of links of interest to CGSC (Fig. 5b) and could be expanded if researchers and database administrators working with other organisms wished to curate entries for their species database. XLocus can be examined at URL http://cgsc.biology.yale.edu/xlocus.html. For example, under Simple Retrieval, enter the Name: icd%, or after selecting Complex Queries, select a Species or Database or Relation or enter a gene symbol (with wildcard % appended) in the Object box. The Help option in various fields will aid users in formulating queries.
FIGURE 5 (Facing page) Retrieval of the Site form for the gene *icd* (top center) from the query (shown in the upper left box) Product = isocitrate dehydrogenase leads to internal hypertext links (indicated by underlining on the retrieved record) to Mutations (below), Strains carrying the alleles (arrow right), and Gene Product (top right), and to external hypertext links to GenBank (far left), Bairoch's Enzyme DB, and SwissProt, with its link to Medline (all upper right), and to similar loci in databases for other organisms via XLocus, shown on this page. (This page) From Site form in the External DB Keys field through XLocus in panel a to similar loci in yeast, maize, human, and *Drosophila* databases. Specific information about these databases and their developers and published references can be found by following the hypertext links to their WWW home pages.

FURTHER INFORMATION

For further information about accessing or using these databases, contact berlyn@cgsc.biology.yale.edu or Mary Berlyn at the address shown in the contributors list at the front of this book.

LITERATURE CITED