Computer System for Maintaining Records on Strains at the *Salmonella* Genetic Stock Centre

ANDREW HESSEL AND KENNETH E. SANDERSON

136

INTRODUCTION

Data management is becoming increasingly important within the laboratory for routine use and legal accountability. For the bacterial geneticist, careful strain maintenance and data management practices are required. Computer-readable files offer many advantages, but few software applications are available for the laboratory market, which is relatively small. This chapter describes the design of the database used by the *Salmonella* Genetic Stock Centre (SGSC) to maintain its collection. In addition, other options for computerization of strain information management are presented.

USE OF COMPUTERS IN STRAIN MAINTENANCE

Any bacterial genetics laboratory generates many hundreds of strains, and efficient techniques of strain handling and data management are thus essential (5). Data management is even more important in the operation of a stock center because of the large number of strains and the need for consistent, accurate, complete strain records. Furthermore, most stock centers normally have to keep records on strain distribution and personal communications pertaining to the acquisition, requesting, or shipping of strains, and thus the number of data statements to be managed may exceed the number of strains by an order of magnitude or more (5).

Although well-maintained and carefully indexed hard-copy records may be adequate for smaller collections, once the number of strains surpasses a few hundred, the use of computers with appropriate software is essential. Computers provide several advantages: rapid searching, updating, and sorting of data; automation of routine clerical tasks; duplication of data records; access to all records from a single site (the computer or computer terminal); and multiple-user access to a single master file (networking). All major stock centers and many research laboratories now maintain their strain data with computers.

Hardware Requirements

Before an electronic data management system can be implemented, the type of computer system to be used must be selected. In the past, many stock centers used mainframes. Microcomputers now have so many advantages that they are widely used, and they are the only type discussed here. Typically, microcomputers with DOS (or Windows), Macintosh, or UNIX operating systems are used. Each system has certain advantages or disadvantages. The improving convertibility between platforms is making the selection of appropriate microcomputer hardware somewhat less important than it once was.

Software

Role of Strain Management Software.

The second basic decision when an electronic data management system is implemented concerns the selection and availability of appropriate software. The primary roles of software in management of data on

strains are to allow searches for strains with desired genetic or other characteristics and to retrieve associated information, such as the source of the strain or a reference about the strain. Larger collections, such as genetic stock centers, may require other software functions for genetic or physical map construction, organization of gene or gene product information, and automation of clerical and other reporting tasks.

Program Types. Typically, strain management software may be divided into three types.

(i) Word processing packages. The simplest method of maintaining strain data is the construction of a table using one of the word processing software packages available for microcomputers. Data in the table may include descriptions of strain properties, references, and locations of the strains. Text-based searches for strain numbers or genotypes are generally possible. Tabular data is nonindexed; thus, Boolean searches (AND OR NOT) cannot be performed. Tabular data may be exported in delimited forms suitable for import into more sophisticated databases. The generation of preformatted information sheets or letters is possible with most word processing packages.

Text-searching utilities that efficiently scan word processor text files are available. Rapid search speed and compact program size are major advantages of such utilities. A useful example of such a utility is StrainFinder, available for MacIntosh computers from Bioware (Table 1). StrainFinder can swiftly search text-based genotypic information while circumventing difficulties associated with double-colon insertional notation (normally overlooked by word processors as nontextual characters). StrainFinder also allows complex searches to be performed by providing Boolean functions and truncation symbols (wild cards) that can be incorporated into the search parameters.

(ii) Commercial strain management programs. A second type of software is a dedicated strain management program either constructed by a commercial vendor or distributed widely as shareware. Only one English language commercial strain management package designed to operate on a personal computer platform is currently available; it is called Gene-Tracker (Table 1). Several freeware and shareware software packages have been described by Sanderson and Ziegler (5); some are included in Table 1, but they may be unsuitable for modern computer platforms.

Commercial or shareware programs may have limitations. Most programs are compiled from source code into a nonmodifiable form prior to distribution; thus, the end user cannot change the program. Modifications normally require consultation with the vendor or distributor in order to correct "bugs," incorporate user improvement suggestions, add features, or utilize new hardware or operating system enhancements. Thus, if a commercial software package is desired, care should be taken to select an established vendor with a good reputation for support.

(iii) Independently developed strain management programs. Because so few commercial strain management systems are available, many laboratories have designed their own software, often using database generation applications such as dBASE, FoxPro, 4D, or FileMaker. (These applications are computer software packages that allow the user to create instructional code that operates within the application environment; in essence, they are programs for building programs.) Database applications differ in flexibility, cost, ease of use, and programming skills required to construct a database. Most will support the relatively simple databases required for strain data management. Independent database development results in database designs that are widely variable, reflecting the different data requirements of the laboratories that generated the software. For example, at the *Bacillus* Genetic Stock Center at Ohio State University, the database called STOX, written and maintained by Daniel Ziegler, maintains records for about 1,500 strains on an IBM-compatible computer (D. Ziegler, personal communication) (Table 1). In contrast, the *E. coli* Genetic Stock Center has constructed a sophisticated database that links strain information with genetic, biochemical, and protein and nucleotide sequence data (1) (Table 1). This database provides excellent support for a large stock center, but the present cost of software support for the Sybase database application and the use of a UNIX-based platform might preclude the use of similar systems by smaller laboratories. Of course, powerful databases can still be constructed by using less

expensive software on smaller platforms, as demonstrated by the *E. coli* genomic databases Colibri (3) and Genescape (2), both of which operate on a Macintosh platform.

TABLE 1 Selected programs used for the maintenance of bacterial strain information

Program name	Type of program	Distribution[a]	Source	Platforms and requirements[b]	Comment
Various	Word processor or text editor	Commercial or shareware	Various	Variable	Low-cost method of maintaining data; nonindexed; limited flexibility
StrainFinder	Text search utility	Commercial; $99	W. Kibbe, Bioware, 1144 Sherman Ave., Suite 100, Evanston, IL 60202; (708) 869-5626; W.kibbe@GEnie.GEIS.com	Macintosh SE or higher, System 7.0, 4 MB	Fast searches; intuitive interface; supports complex Boolean searches; prints search results
CLONES	Strain and plasmid management	Shareware; $10	Hal B. Jensen, Yale University School of Medicine, New Haven, Conn.	IBM (DOS 2.1 or higher)	Suitable only if newer hardware is not available
STOX	Strain and plasmid management	Noncommercial, independent	D. Ziegler, *Bacillus* Genetic Stock Center, Ohio State University, Columbus, OH 43210; (614) 292-5550; dzeigler@magnus.acs.ohiostate.edu	IBM or compatible	Source code only provided; requires compiler
Gene-Tracker	Strain and plasmid management	Commercial; $800/$300(academic)	J. Rothfield, P.O. Box 394, San Rafael, CA 94915; (510) 525-5742	Macintosh, SE or higher, System 7.0, 2 MB; Windows 3.1, 8 MB	Full-function database; can be customized by vendor
DataMinder	Laboratory utilities and organizer	Noncommercial, independent; available by ftp at bioinformatics.weizmann.ac.il/pub/software/mac	Karen Usden, NIH, Bethesda, MD 20892-0830; ku@helix.nih.gov	Macintosh, SE or higher, System 6.0.5 (min), 2 MB	Limited strain management functions; many other useful features; data are exportable
E. coli Genetic Stock Center Database	Strain and genetic data management and interrelated functions	Noncommercial, independent; not intended for general distribution	M. Berlyn, *E. coli* Genetic Stock Center, Yale Univ., P.O. Box 6666, New Haven, CT 06511-7444; berlyn@cgsc.biology.yale.edu	Sun (UNIX)	Sophisticated, powerful system; considerable development and support costs
SGSC Strain Information Management System	Strain and plasmid management	Noncommercial, independent; free (requires FileMaker software)	*Salmonella* Genetic Stock Centre, University of Calgary, Calgary, AB Canada T2N 1N4; (403) 220-3572; kesander@acs.ucalgary.ca	Macintosh, SE or higher, System 6 (System 7 recommended), 4 MB	See text.

[a]All prices are given in U.S. dollars.
[b]MB, megabytes.

THE SGSC DATABASE

The main functions required of a database at the SGSC are (i) maintaining strain data in a standardized and easily searchable format; (ii) allowing searches on a variety of criteria to find strains to satisfy the needs of requestors; and (iii) automating routine clerical duties required in the daily functions of the stock center. Efforts

to integrate strain data maintained by the SGSC with that of the *Salmonella* genetic map (4) (see chapter 110) have not yet been completed. An earlier strain management database designed at the SGSC by an independent consultant using the application program dBASEIII with a DOS-based machine was briefly discussed earlier (5). In order to allow the use of Macintosh computers and because of some inflexibilities in the earlier system, a new system was recently developed; this is described below.

FileMaker Pro (Version 2.0v3, Claris Corp., Santa Clara, Calif.) (hereafter referred to as FileMaker) was selected as the database application to provide the operating environment for the new SGSC database. FileMaker has these advantages: it is a widely used, inexpensive program (about $150 US); it simplifies the construction of powerful databases; it is available for both Macintosh computers and IBM-compatible equipment running Windows, and allows data files to be shared between the two computers; it allows direct importation of many different file formats (including the one used by the original SGSC database information files); and it requires minimal computer experience to operate. The FileMaker application includes good documentation, drawing and coloring tools for the development of graphical layouts, a tutorial, help files, and spelling correction software.

Database Construction

Database construction in the FileMaker environment is managed by the use of graphical presentation layouts called "templates." Following the definition of the names and types of data fields to be included in the database template, the user builds layouts to format the information for screen presentation or printing. Multiple layouts may be designed to utilize a single set of data fields; for example, strain data could be presented on the computer screen with one layout but reorganized with a different layout for print output.

User-defined operations to be performed on the data are specified by the construction of scripts, which are series of data or file manipulation steps. No formal programming experience is required to write a script. All possible operations are presented in the form of a menu list; the user assembles a script by selecting the required steps and arranging them in a logical sequence. Although each step is clearly defined in the user's guide supplied with the application, most script steps can be understood intuitively. Scripts can be executed by assigning them to a graphical button in a layout or to a "quick key." (A quick key is a defined keystroke, such as depressing the and 1 keys simultaneously, that is user defined during script creation.)

Changes made to the number or type of fields or to layouts and scripts are immediately implemented; there is no need to recompile the source code into a new executable file. The ability to rapidly and conveniently alter the database design prevents program obsolescence and should allow databases to remain useful for longer periods.

Data can be ported to or from the database in a variety of other file formats, and viewing of the data, or modification of it, can be restricted by user password access. Database layouts and information can be shared across and between MacIntosh and IBM networks (although some additional hardware may be required to bridge the different platforms); this allows multiple computers to access and maintain a single master data file. Data recovery following a system crash or power failure is automatic and reliable. Large database files can be maintained in a compressed format by using a built-in compression utility, reducing disk storage requirements by as much as 50%.

Features of SGSC Data Files

The complete SGSC database, collectively described as the Strain Information Management System, was constructed by using FileMaker. The complete system consists of eight smaller, interlinked database templates (Fig. 1). Each of the templates is a stand-alone module that performs a specific role; partitioning the databases in this manner keeps unrelated functions and information separate and simplifies data maintenance. Keeping separate template files also minimizes the loss of data or database function if one of the template files becomes corrupted.

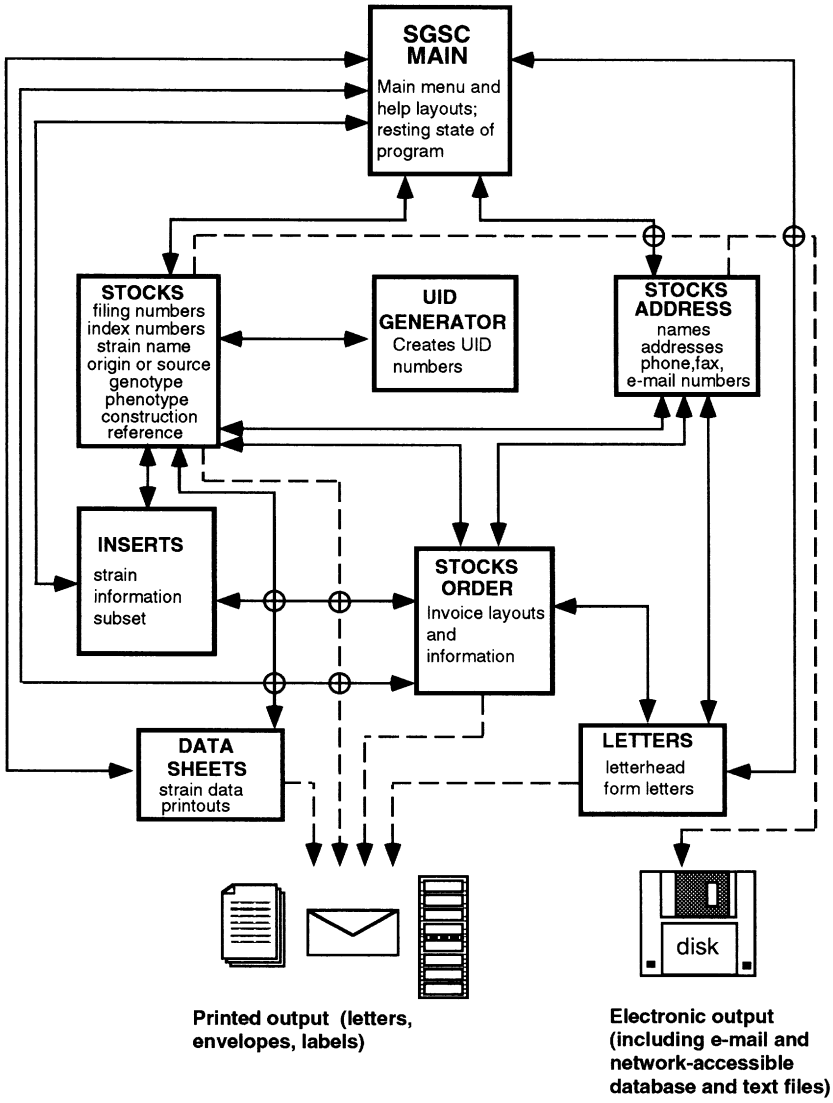

FIGURE 1 Schematic of SGSC Strain Information Management System design. The eight database templates are depicted as boxes; database names are in boldface capital letters. Partial information on what is contained within each template or the primary function of each template is listed inside each box; complete template functions are described in the text. User-controlled navigation or preprogrammed information exchange between linked database templates is depicted by solid arrows. All templates, except UID GENERATOR, can be accessed from the SGSC MAIN menu system. Information output is depicted as dashed arrows. Information contained within the templates can be printed or exported directly to a variety of data file formats (see text).

Most of the SGSC templates contain multiple layouts. The main layout typically displays information contained in the database and also several button choices grouped into a command menu; pressing on a button with the mouse-directed pointer initiates either a scripted data action or program navigation. All command menus include a selection that allows the user to return to the main menu (the program resting state). Layouts for printing reports, displaying user warnings, and giving help information are also included.

Template Descriptions

SGSC MAIN. The SGSC MAIN template consists of several graphical layouts that display program functions as menus. Navigation options are presented as buttons in each menu. “Clicking” on a button with the mouse-controlled pointer directs the program to open the desired database and execute the requested function. This template is for user convenience only and has no information fields or records to be damaged if a power failure occurs. Thus, to avoid potential data loss, users should return to the SGSC MAIN template when they are not accessing other strain management functions.

STOCKS. The STOCKS template contains the majority of the information on the bacterial strains at the SGSC. The following information is stored within this database: genotype, phenotype, species, storage site, key gene, construction information, and strain references. All data fields are fully indexed except for the genotype, phenotype, and reference fields; information within these fields is delimited only by spaces. Although nonindexed genomic information limits the searches that can be performed and prevents the use of individual genetic characteristics as pointers to other information, it simplies the maintenance and visualization of strain data. (Other fully relational databases to satisfy more sophisticated searches are currently under construction.)

Each strain entered into the STOCKS template is assigned a unique identifier (UID) number. This numerical code can be used as an index to information about the strain contained in the other database templates.

The strain information layout within the STOCKS template allows visualization of strain records (browsing), searching for strain information, addition of new strain records to the database, printing of strain information, and navigation to other functions or databases. An example of the STOCKS strain information graphical layout is presented in Fig. 2.

STOCKS ADDRESS. The STOCKS ADDRESS template contains fields for the names, addresses, telephone and fax numbers, and e-mail addresses of the researchers who have either ordered from or contributed to the SGSC collection. The researcher identification (RID) numerical pointer is used to link the researcher with strains provided to the SGSC. The address template is also accessed by the LETTERS and STOCKS ORDER templates (Fig. 1).

STOCKS ORDER. The STOCKS ORDER template is used to process orders received by the stock center. The graphical layout is presented in . 3. The following steps are performed to fulfill a request: the recipient of the order is found in the STOCKS ADDRESS template by using the Find Researcher button; a new invoice is created; a strain to be sent is entered into the strain field; additional strains are appended to the order with the Add Strain button; invoices, data sheets, and covering letters may then be produced by selecting the appropriate buttons. The STOCKS ORDER template allows statistics to be generated, for any date range, on the number of strains sent, the strains ordered by any researcher, the average days required to fill an order, and the number of times any strain has been sent.

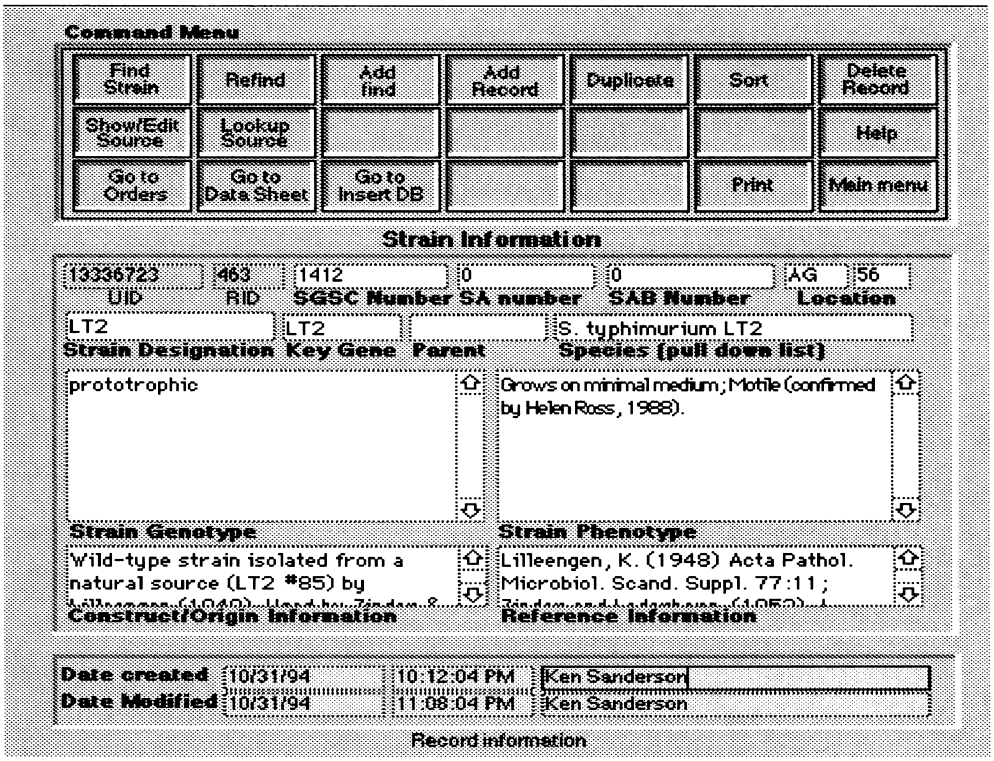

FIGURE 2 Strain Information Management System layout STOCKS template. Strain information contained in each database record is presented in a graphical layout composed of information-containing fields (boxed text) and a command menu. The command menu offers functional or navigational options in button form. Fields that may contain more information than can be displayed within the field dimensions (e.g., Construct/Origin field) include arrowed controls at right-hand edge of field that allow the user to scroll through the text contained within the field. Selecting a command menu button with a mouse-controlled pointer executes a programmed action, or script (see text). All command menus allow navigation to the SGSC MENU database by using the Main Menu button or to database-specific help information. Filing numbers, such as those contained in the UID and RID fields, are used to index the strain record to information in other linked templates (Fig. 1). To maintain data entry consistency, species information is entered from a pull-listing of strains. Records are retrieved in searches by matching information entered into one or more fields.

DATASHEETS. The DATASHEETS template prints strain data information sheets without generating permanent records (such as those produced by the STOCKS ORDER template).

LETTERS. The LETTERS template can be used to create either original correspondence or form letters on laboratory letterhead. Researcher address data are automatically retrieved from the STOCKS ADDRESS template.

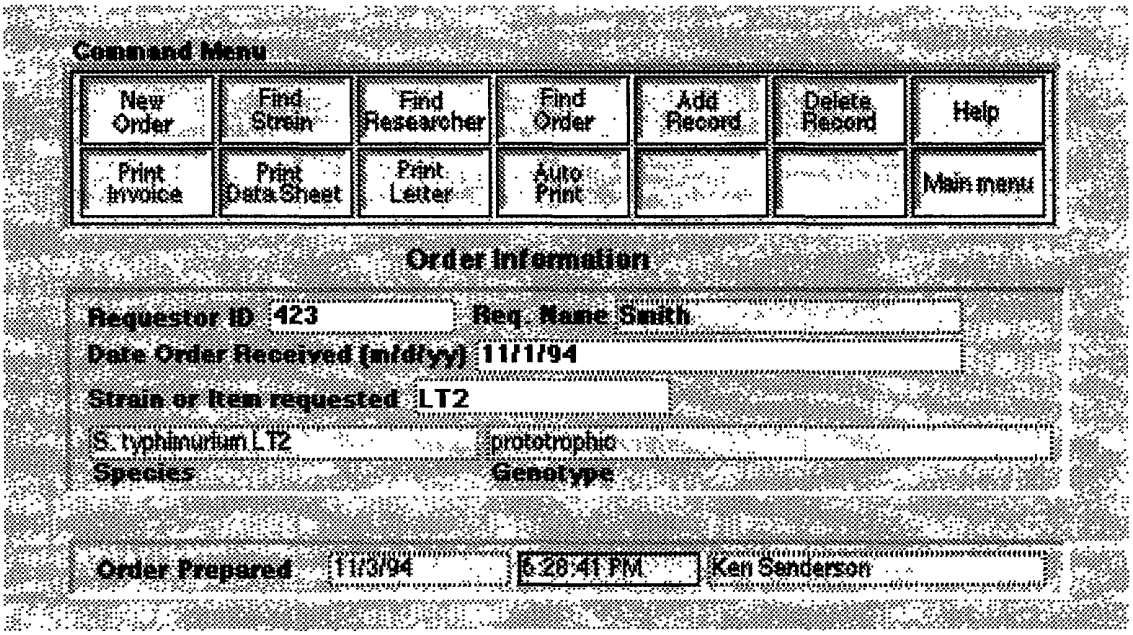

FIGURE 3 Order entry layout of the STOCKS ORDER template. Database records are created during order processing. The researcher who requested the strain is identified by the RID index number; this number is obtained by navigation to the STOCKS ADDRESS template by using the Find Reseacher button. Strains to be included in the order are entered into the strain name field; information on each entered strain is automatically extracted from the STOCKS data records. Strain searches may be performed by selecting the Find Strains button in the command menu; this allows navigation to the STOCKS database template. Following entry of all strains to be included in an order, the strain information sheet, invoice, and form letter to be included with the shipment can be automatically printed by selecting the appropriate buttons from the command menu.

INSERTS. The optional file INSERTS is used to maintain a subset of information obtained from the STOCKS template on strains with characterized insertions of transposons into known genes or chromosomal regions. Information (such as strain name, SGSC number, and laboratory of origin) was copied to the INSERTS template data fields from the STOCKS template on strains with transposon (such as Tn*10*) insertions. Additional information was then added to each strain record, including map minute, location of the insertion, antibiotic resistance associated with the insertion, and known linkage of the insertion with other insertions or mutant alleles. This template will not be required when the genotypic data maintained in the STOCKS database are reorganized to be completely indexed.

UID GENERATOR. The single-function UID GENERATOR template is accessed only when a new strain record is created. It contains a calculation-type field that is used to generate a unique numerical pointer to be included with each new strain record.

Data Searches and Limitations

Strain searches are performed by pressing the Find Strains button in the STOCKS template command menu (Fig. 2), entering data into the appropriate field (for example, the strain name or key gene field), and then pressing the keyboard Return key. The use of numerical operators (e.g., < or >), wild cards, (e.g., * [finds one or more characters] or @ [finds any single character]), and other special functions can be included in the search parameters. Logical AND searches are performed by search criteria added to several fields.

While the nonindexing of genotype information does not complicate simple searches, a difficulty occurs in attempting to find insertional mutations. Transposon insertions are denoted with a double colon (::), while punctuation within data fields is ignored by the FileMaker program. Thus, a search for *proC@@@*::Tn*10* (gene *proC* followed by a three-digit allele number and then a Tn*10* insertion) will locate records matching the criteria *proC@@@ AND* Tn*10* in any order or combination within the genotype field. Utilizing this search string, records with the genotype entries *proC111*::Tn*10* and *proC345 thr-123*::Tn*10* are equally likely to be found.

Logical OR searches can also be performed on strain data but only as two consecutive find requests. (Find requests are stacked by using the Add Find button included in the STOCKS command menu [Fig. 2].) Each individual find request will be subject to the search limitations previously described.

In general, searches of the ca. 5,000 SGSC records are performed in less than 1 s, although mutiple-request or wild-card text searches take longer. Strain data can be exported into a file suitable for use with rapid text-scanning utilitity programs, such as StrainFinder, which will efficiently search for insertions of specific transposons in specific genes.

Distribution of SGSC Data and Databases

The Strain Information Management System described here may have applications in other laboratories. A generic version of this system, stripped of SGSC address headers and other laboratory-specific information, is available from the SGSC as a database clone (database without records). The cloned database can be easily modified by other users to suit their own information-handling requirements or personal tastes in graphical data presentation. Modification of the original database design is encouraged, although secondary distribution of the program is not. The user must have the FileMaker application program for either Macintosh or IBM Windows in order to use and modify the SGSC databases.

Researchers who wish to view SGSC strain information records but do not require a complete Strain Information Management System have two options. First, a limited version of the SGSC MAIN, STOCKS, and INSERTS database templates is available from the SGSC as the SGSC Strain Information Viewing System. These files represent a subset of information and functions provided by the complete information management package. Records may be searched, displayed, and printed, but these records and the database designs cannot be altered. The FileMaker application is still required to access these databases. Second, a tab-delimited text file of strain information, searchable with a word processor or text-searching utility and importable by most database applications, is available. The FileMaker application is not required in order to utilize this file or to modify its contents, and it is available in both Macintosh and IBM diskette formats.

The SGSC database systems and the delimited text file of strain information can be obtained at no charge from the SGSC. Whenever possible, a request should be accompanied by two high-density diskettes to help defray material costs. Alternatively, the database files (maintained as binhexed Stuffit Self-Extracting Archives) and instructions on their use are available from the University of Calgary via Internet, using the gopher menu system. Interested parties should navigate to the University of Calgary gopher and then select, in order, Faculty and Departments Information, Department of Biological Sciences, and Salmonella Genetic Stock Centre. The files may also be retrieved from the SGSC World Wide Web page (http://asc. ucalgary.ca/~kesander). Comments and difficulties with the databases or their acquisition or requests for technical assistance should be directed to kesander@acs.ucalgary.ca.

ACKNOWLEDGMENTS

The generous contribution of strains and data by the many researchers who have donated materials to the SGSC is gratefully acknowledged. We were supported by an Operating and an Infrastructure grant from the Natural Science and Engineering Research Council of Canada and by grant R01AI34829 from the National Institute of Allergy and Infectious Diseases while this report was being prepared.

LITERATURE CITED

1. **Berlyn, M. B., and S. Letovsky.** 1992. Genome-related datasets within the *E. coli* Genetic Stock Center database. *Nucleic Acids Res.* **20:**6143–6151.
2. **Bouffard, G., J. Ostell, and K. E. Rudd.** 1992. Genescape: a relational database of *Escherichia coli* genomic map data for Macintosh computers. *Comput. Appl. Biosci.* **8:**563–567.
3. **Medigue, C., A. Viari, A. Henaut, and A. Danchin.** 1993. Colibri: a functional data base for the *Escherichia coli* genome. *Microbiol. Rev.* **57:**623–654.
4. **Sanderson, K. E., and J. R. Roth.** 1988. Linkage map of *Salmonella typhimurium,* edition VII. *Microbiol. Rev.* **52:**485–532.
5. **Sanderson, K. E., and D. R. Ziegler.** 1991. Storing, shipping and maintaining records on bacterial strains. *Methods Enzymol.* **204:**248–264.