MICROBIOLOGICAL REVIEWS, Sept. 1993, p. 623–654
0146-0749/93/030623-32$02.00/0
Copyright © 1993, American Society for Microbiology
Vol. 57, No. 3

Colibri: a Functional Data Base for the *Escherichia coli* Genome

CLAUDINE MÉDIGUE,[1] ALAIN VIARI,[2] ALAIN HÉNAUT,[3] AND ANTOINE DANCHIN[1]*

Régulation de l'Expression Génétique, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15,[1] Laboratoire de Physique et Chimie Biomoléculaire, Université de Paris VI et Institut Curie, 75231 Paris Cedex 05,[2] and Centre de Génétique Moléculaire du Centre National de la Recherche Scientifique, 91198 Gif sur Yvette Cedex,[3] France

INTRODUCTION

One of the goals of biologists is to understand the basic mechanisms that govern the building up, propagation, and evolution of living organisms. Sequencing programs are designed to decipher messages corresponding to genetic programs of procaryotic and eucaryotic model organisms. A preliminary step is to complete the total genome DNA sequence of the organisms. National and international libraries have been collecting information on nucleotide and protein sequences for many years. This is a major contribution to the accessibility of biological knowledge, but the information thus collected cannot be directly treated for proper handling by specific software. It was therefore necessary to evaluate the feasibility of constructing a data base from *Escherichia coli* by using the data present in the banks. An expected consequence is that exploration of the "genomic text" should result in the discovery of rules that govern its organization and operation, namely to estimate its consistency. Computing is then required for three distinct but interrelated operations: data acquisition, data exploitation (i.e., extraction and interpretation of sequence information), and finally management of biological data derived from the two former operations. Setting up a consistent informatic system at this level requires the integration, within a single environment, of (i) a specialized data base, (ii) sequence analysis software, and (iii) biological knowledge. Advances in data acquisition techniques, as well as the proliferation of computer-assisted tools, should lead to a sophisticated study of the genomic text and the exploration of hypotheses on regulation of gene expression and on gene function and evolution.

As a preliminary step for determining such an environment, we have built up a specialized data base by using biological data currently available for the study of *E. coli*. The genomic molecular structure of this microorganism has already been studied intensively for many years and today provides a richer set of data than that of any other known living organisms. Since 1976, when Taylor and Thoman identified and positioned 99 genetic loci on the *E. coli* chromosome map, eight further editions of the linkage map have been published. In the most recent map, Bachmann positioned more than 1,400 loci, representing about one-third of the total gene content of *E. coli* (1). A complete restriction map of the chromosome of strain W3110 was constructed by Kohara et al. (16). Since then, several programs have been developed to correlate the DNA sequence data directly with the physical map of *E. coli*. This software is based either on a method of restriction pattern alignment (23, 26) or on the comparison of the length of each restriction fragment (19). Since 1989, Kröger et al. have been compiling *E. coli* nucleic acid sequences from the GenBank and EMBL data libraries. In the last update 38.5% of the

* Corresponding author.

entire *E. coli* chromosome was described (17). In a similar way, Rudd et al. (27) have developed a software (including both data files and application programs) for collecting, aligning, and displaying *E. coli* genetic map (1), restriction map (16), and DNA sequences obtained in different laboratories after removing redundancy.

Independently, we started collecting *E. coli* DNA sequences from the EMBL data library together with miscellaneous data obtained from other sources. Our main purpose was to build up a set of clean, consistent, and nonredundant data and to extract, in the best conditions, information about *E. coli* nucleic acid sequences. In an attempt to evaluate the consistency of the sequences generated by the laboratories working on *E. coli*, we found that the amount of polymorphism at the nucleotide level was small, justifying the aggregation of such data to generate a "patchwork" of the *E. coli* chromosome sequence (19). As our work on this genome progressed, it became necessary to define a structuring model for the *E. coli* biological data, allowing easy consultation as well as multicriteria searches of any information whether it originates from experimental work or from computer analysis of DNA sequences. In contrast with Rudd et al., who use a software that manages independent data files (27), we used a relational data base management system to conceive a specialized data base for the *E. coli* genome. This data base, called Colibri, is presented in this paper. We describe the organization of data, the structure of the data base, and the environment of consultation and interrogation. Graphic representation of the data stored in Colibri shows that the modeling of biological knowledge makes the extraction of new information significantly easier. To date, the data base holds more than 2,000 kbp of *E. coli* DNA sequences, i.e., 50% of the total genome. An update, including a procedure aiming at the construction of contigs without the help of the restriction map, will be presented elsewhere. We have defined in the present work a way of modeling procaryotic biological data. This model is presently being used to build up a specialized data base for the *Bacillus subtilis* genome.

SOFTWARE AND DATA

Until now, biologists have been using relational data base management systems to organize DNA sequence data; this is how the GenBank data base (5) and the PIR protein sequence data base (2) are organized. This choice is made mainly because relational data models support a very simple data structure (i.e., in tabular form) easily understandable by those who are not computer professionals. Besides, these systems provide ad hoc query languages which are simple to use. The relational data base system used as part of this work and the various *E. coli* biological data that we had to organize in the base is briefly presented below.

Data Base Management System

The data base Colibri has been developed on a Macintosh IIcx by using a commercial relational data base management system, 4th Dimension (4D). This software is user friendly and greatly facilitates the task of prototyping and refining an operational data base. 4D allows description and handling of data as file forms (i.e., tables in the standard relational model). A 4D file is made up of fields of alphanumerical, real, integer, boolean, etc., types (i.e., the attributes of the relation scheme). Basic elements of a file are the cards or the records (i.e., the t-tuples). An advantage of 4D is that it is less restrictive than the standard relational model since it offers the capacity to manage image or text data types. Accordingly, 4D is very well suited to model biological objects such as sequences, gels electrophoresis patterns, genetic maps, and physical maps.

The relationships between structured data, i.e., between the 4D files, can be of different type. For instance, the biological object "sequence" is characterized by simple attributes such as its name, its length, and its accession number in the DataBank, but also by more complex properties such as the description of its interesting biological properties (called features). Similarly, the object "gene" is characterized by its biological function (specified by key words), among other properties. The relationship between objects "sequence" and "feature" corresponds to a $1 \leftrightarrow n$ Mapping (mapping is used here in the mathematical sense. To avoid confusion with mapping in the sense of genetics, we use a capital M); i.e., one sequence is characterized by several biological sites, and a particular site is specific to only one sequence (Fig. 1A). On the other hand, the relationship between objects "gene" and "key word" corresponds to the Mapping $n \leftrightarrow m$, i.e., one gene is characterized by several key words, and a particular key word is generally specific to several sequences (Fig. 1B). The structure generator of 4D makes it possible to translate an injective Mapping ($1 \leftrightarrow n$) by tracing a link between two files with the mouse. A link associates N cards of the first file with one card of the second file (Fig. 1A). To establish an $n \leftrightarrow m$ Mapping between two files, it is necessary to create a "buffer file" that presents a $1 \leftrightarrow n$ Mapping with the first file and with the second file. In Fig. 1B the 4D file [Keywords] is a buffer file. This data organization (several linked files) eliminates redundancy since data in files are recorded independently and only once.

The exploitation of a data base conceived with 4D is realized by means of a personalized interface. It consists mainly of various layouts presenting data on the screen, dialogues, but also procedures connected to buttons or to menus. The generator of a 4D application associates the faculties of a fourth-generation language with those of Pascal or C (mathematical function, recursivity, etc.). Data bases developed with the update 4.1 of the software can also be compiled, thus allowing fast running of the procedures. Besides, 4D is completely open to the outside as far as it is possible to export or import data, integrate external procedures (written in Pascal or C) into a data base, but also to link up a base with other relational data base management systems.

Data Available from the *E. coli* Chromosome

The *E. coli* genome consists of a long, circular, supercoiled DNA molecule with 4.7×10^6 bp. This single chromosome can be represented in one of three different ways: the genetic map, the restriction map, and the DNA sequences.

Genetic map. The genetic map is represented by a collection of genes that have been identified by using mutant phenotypes and ordered by using information from genetic crosses. The *E. coli* genetic map has been compiled by B. Bachmann; in the last update, more than 1,400 genes were identified and mapped on the chromosome (1). The genetic map is used to collect experimental information about genes by ordering their chromosomal positions. It was thus important to organize the corresponding knowledge in Colibri.

Restriction map. The restriction map is represented by a

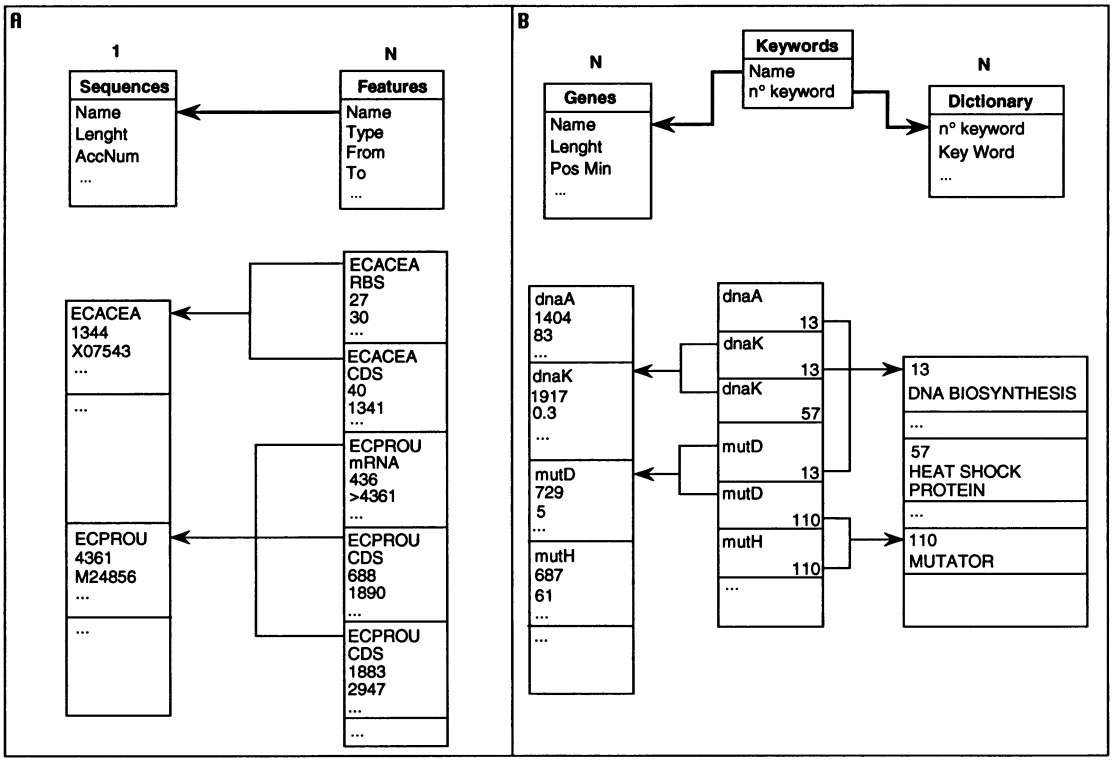

FIG. 1. Relationships between structured data in a 4D data base. In the upper part of the figure we have represented several 4D files [Sequences], [Features], [Genes] etc., comprising fields of different types (alphanumeric type for the fields Name or AccNum, integer for the field Length, etc.). The lower part of this figure shows examples of records for each of these files (for example, in the first record of the [Sequences] file, the field Name is equal to "ECACEA," the field Length is equal to 1,344 and the field AccNum is equal to "X07543"). (A) Illustration of the $1 \leftrightarrow n$ relationship between the files [Sequences] and [Features] (see text). (B) Illustration of the $n \leftrightarrow n$ relationship between the files [Genes] and [Dictionary]. The file [Keywords] is a buffer file (see text).

collection of endonuclease sites distributed along the DNA molecule. This generates DNA fragments that can be ordered by using molecular cloning, gel electrophoresis, and DNA hybridization techniques. Such a map is useful for the cloning and sequencing of genes. The initial *E. coli* physical map was established by Kohara et al. (16). It contains information about the distribution of patterns that are recognized by eight restriction enzymes: *Bam*HI, *Hin*dIII, *Eco*RI, *Eco*RV, *Bgl*I, *Kpn*I, *Pst*I, and *Pvu*II (7,108 restriction sites in total). However, the standard Kohara map included seven known gaps that were subsequently filled by Knott et al. (14, 15). In parallel, local restriction maps can be generated from known *E. coli* nucleic acid sequence. This new information can be used to correct and supplement the initial Kohara restriction map (see below).

***E. coli* DNA sequences.** The *E. coli* DNA sequence map is represented by an ordered collection of nucleotides represented as one of four letters A, T, G, or C. DNA sequencing allows precise length determination of any region of a genome and reveals signals in DNA such as coding sequences and control regions. The *E. coli* sequence map is still incomplete; the last release of the EMBL data library we discuss here contained 2149 entries that would correspond to about 56% of the entire *E. coli* genome, if they were not redundant. In fact, over 40% of the *E. coli* chromosome has been sequenced to date as a collective but uncoordinated effort (17).

In the literature these data are not related to each other. Therefore, as the work on *E. coli* genome progresses, the main scope of our work is to integrate all three types of information into a single map to make consultation, utilization, and updating of data easier. It was therefore necessary to define a representation of these data with 4D management system.

LOGICAL STRUCTURE OF THE DATA BASE

The conception of a data base first requires an analysis of the users' needs, i.e., definition of the way of viewing and recovering information. This consists of describing very precisely all data to be modeled, finding the relevant data, and characterizing their intrinsic relationships. Another aspect, closely linked to the first, consists of determining the expected representation and processing. The structure of the data depends on such an analysis and consequently on the efficiency of the data base (capacities and performances) and the ease with which it can be modified. We show in Fig. 2 the logical structure of the data base we have developed.

Data from Public Libraries

Information describing each DNA sequence collected in the EMBL format is organized in several linked files (Fig. 2). The data are used mainly as reference documentation, since users of the data base are supposed to have access to original EMBL information. Thus, the [EMBL] file contains fields corresponding to the ID (identifier), AC (accession number), DE (description), CC (comments), and SQ (sequence) lines

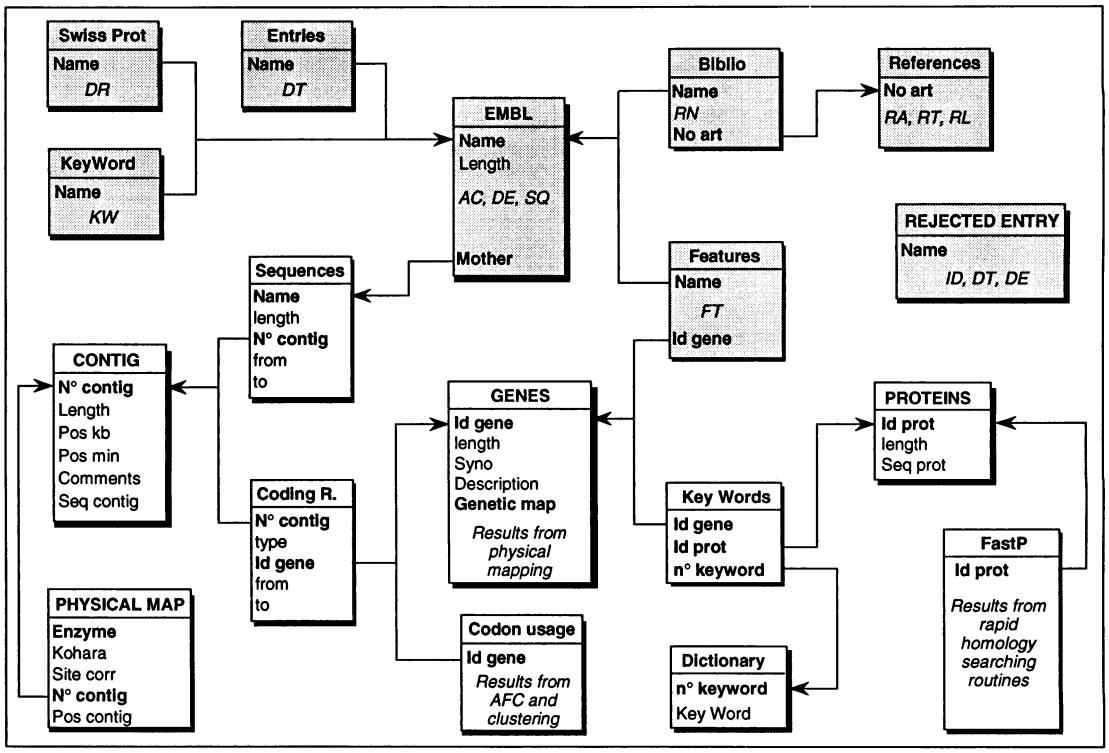

FIG. 2. Representation of the logical structure of Colibri. The various *E. coli* biological data are organized in the different 4D files represented in this figure: data from public libraries (displayed as shaded files), contigs, physical map and genetic map, and proteins, but also data corresponding to results of various software analyses (shown in italics). The relationships between the 4D files are represented by arrows linking two indexed fields (shown in boldface type). AFC, factorial correspondence analysis.

of the EMBL data file format (10). To avoid redundancy (see "Generation of Contigs"), data corresponding to the DR (access to the data bank Swiss Prot), DT (creation or modification of the entry), KW (key words), FT (Features table), and finally RP, RA, RT, and RL (bibliographic references) lines have been structured in several independent 4D files (Fig. 2). The relationship between the [EMBL] file and others follows a $1 \leftrightarrow n$ Mapping (for instance with the file [Features]) or an $n \leftrightarrow m$ Mapping (for instance with the file [References]) (see above). Thus, all data describing a single EMBL entry can be recovered by using the links described in Fig. 2. Finally, the [REJECTED_ENTRIES] file contains EMBL entries corresponding to RNA sequences, plasmid or pathogenic *E. coli* strain sequences, and DNA sequences which were not obtained from the *E. coli* K-12 genome but were mistakenly recovered under *E. coli* keywords. Information of each release of the EMBL data bank received in CD_ROM form, are automatically extracted. From every EMBL release, new entries are created and those already existing are modified or sometimes removed. A few sequences directly obtained from individual laboratories are also included. Procedures have thus been built in to update the data base content. For example, with the EMBL release 31 there are 1,460 [EMBL] file cards and 689 [REJECTED_ENTRIES] file cards, both corresponding to the total entries of the data library.

Generation of Contigs

A necessary step in the development of Colibri is related to the detection of duplicates. Two kinds of multiple entries are present in the original EMBL data bank. They may correspond to duplicate sequences of exactly the same gene. These sequences are almost identical. In this case, the choice to be made requires biological expertise and cannot be done automatically. Most of the time, one entry is a fragment entirely present in a long sequence forming another entry. It is easily detected by using the FASTP identity-searching routines (30). Sequences that appear only once are called "Mother" (Fig. 2, [EMBL] file). After elimination of duplicates, we generate a series of longer DNA sequences termed "contigs," when sequences overlap with two or more neighbouring sequences. This represents the nonredundant *E. coli* sequencing information. It is structured in the 4D file [CONTIG]. Links between [Sequences] and [EMBL] files, on the one hand, and [Sequences] and [CONTIG] files, on the other hand, allow the immediate recovery of all EMBL information for each entry of a particular contig (Fig. 2). A procedure has been constructed for checking possible mistakes in the assignment of coding sequences (CDSs). Obviously, this requires verification in original publications in many instances and cannot be performed automatically. Data concerning the CDS localization on contigs and their properties are organized in the 4D file [Coding Region]. When the Colibri data base is being updated, a procedure identifies contigs to be created or modified, genes already existing in the data base (i.e., duplicates), and, when needed, possible mistakes in new coding sequences. A total of 495 overlaps and single DNA sequences are currently recorded in the 4D file [CONTIG].

Genetic Map and Proteins

Information of each sequenced gene is also organized in the data base ([GENES] 4D file [Fig. 2]): its name, synonyms, map position, phenotype, etc. This information is generally extracted from the Bachmann genetic map (1). Other properties of the genes were obtained after performing a statistical analysis of the data ([Codon usage] file; see below) or from the mapping of DNA sequence onto the physical map of the *E. coli* chromosome. The file [PROTEINS], linked to the [GENES] file by means of the buffer file [Key Words] (see above), is generated during a procedure which creates records for the amino acid sequences translated from coding regions. For each protein record stored in the data base, we use the FASTP searching routine (30) and organize results in the 4D file [FastP] (see the section on calling external procedures, below). Rapid searching for a particular gene or protein function is realized through a key words list constructed from those given in the EMBL and Swiss Prot data banks together with those of Bachmann genetic map (1). This preestablished list is thus structured in the [Dictionary] file (Fig. 2). The number of sequenced genes currently recorded in Colibri approaches 1,500. The biological function of 200 of these genes remains unknown.

Physical Map

Finally, data corresponding to the Kohara physical map (16) are also structured in Colibri. The file [PHYSICAL MAP] is made up of fields created to record all positions of the eight restriction enzymes sites. We have developed a program, written in Pascal, that identifies the most likely positions of a DNA sequence on the *E. coli* chromosome by using information about restriction fragments generated with the eight enzymes used by Kohara et al. to map the genome (16). This software, described previously (19), has been used to assign a map position to most of the contigs constructed in Colibri. Appropriate localization of a fragment, correlated to data obtained by classical genetical means, allows us to suggest corrections of the Kohara map. These corrections are performed mainly by adjusting the positions of restriction sites and by adding missing sites to the original physical map. It was found that, taken as a whole, the Kohara map is very accurate, except maybe for the *Pvu*II and *Eco*RV sites, which seem to be context sensitive (20). Since there are more missing sites than extra sites in the last update of our corrected physical map, the most likely explanation is that the method used by Kohara et al. (partial digestions) resulted in a defect in restriction sites. Then, it seems that there is only a low level of polymorphism between the *E. coli* K-12 strains used in the various laboratories involved in DNA sequencing (19). By updating data obtained from the physical mapping (19, 20, 22), together with data on restriction sites, positions of each new contig are organized in several fields of the 4D file [PHYSICAL MAP]. This file is linked to the [CONTIG] file by an $n \leftrightarrow 1$ Mapping (Fig. 2). Our corrected restriction file is available on request to scientists providing electronic mail address. At present, it contains nearly 8,000 restriction sites.

USING COLIBRI

The second step when constructing a data base is related to the representation of the data on the screen. As depicted in Fig. 3, the different ways of representing the *E. coli* chromosome are closely connected to the biologist's mental representation. In the previous sections we have explained how a data base management system such as 4D allows the modeling of these data. Structured information in a data base must express the user's views in the best way, but this physical organization is quite invisible to the user. That is why a data base management system also provides an interface generator. One thus defines a set of layouts presenting data on the screen: this is what can be seen by the user (Fig. 3).

Interfacing Colibri

Several layouts have been developed in the Colibri data base which contain combinations of data from different 4D files (Fig. 2). This environment of data consultation and multicriteria searches provides a palette of tools (buttons, pop-up menus, dialogues, etc.) associated with particular procedures for searches, "navigation" in the data base, and importation of data in ASCII file form. Other 4D procedures are linked to the interface menus and allow particular analysis of the *E. coli* DNA sequences, such as finding the position of restriction sites, performing translation, and searching for similarities in protein data banks. The user can start the data base consultation from any type of data: EMBL entries, contigs, physical map, genes, etc. The relationship defined between the structured data allows direct access to all the related information. For instance, for a set of genes which express a same biological function, the user can very easily find all the corresponding EMBL entries together with the bibliographical references, as well as the genomic environment of those genes (e.g., the contigs they are on, their chromosomal localization, their neighboring genes, their restriction map). Furthermore, by using the drawing unit of 4D (4D Draw), it becomes possible to integrate an interactive graphic representation of the data to the interface of Colibri (see below). Very soon, most of the text information of the data base should be accessible in sensitive areas on the represented drawings of the genetic map, physical map, or DNA sequences and their coding regions.

Examples of *E. coli* Data Representation

As an example we show in Fig. 4.A the layout of data for a recorded contig localized at 83.7 min. The information presented here is length and genomic address (kilobases) of the genomic restriction site (16) that is mapped to the first restriction site of this contig (the orientation of its DNA sequence is identical to the direction of increasing map coordinates). The user can export the DNA sequence as an ASCII file. The map position of the contig (in minutes) is calculated from its genomic address (assuming a length of 4,719.6 kb and 100 min for the *E. coli* chromosome). Information from EMBL entries used to build up the contig are indicated in the EMBL Sequences table, i.e., their name in the data bank, their relative position on the contig, and their orientation with respect to the original one (a complementary strand is indicated by a minus sign). The user has access to all EMBL information of a particular EMBL entry with the "Info EMBL" button. Finally, in the Coding Regions table, we indicate information concerning each gene included in the sequence of the contig (its name, position, and biological function). The Graph button is linked to a 4D procedure that allows drawing positions of the coding regions and the restriction enzyme sites on the contig (Fig.

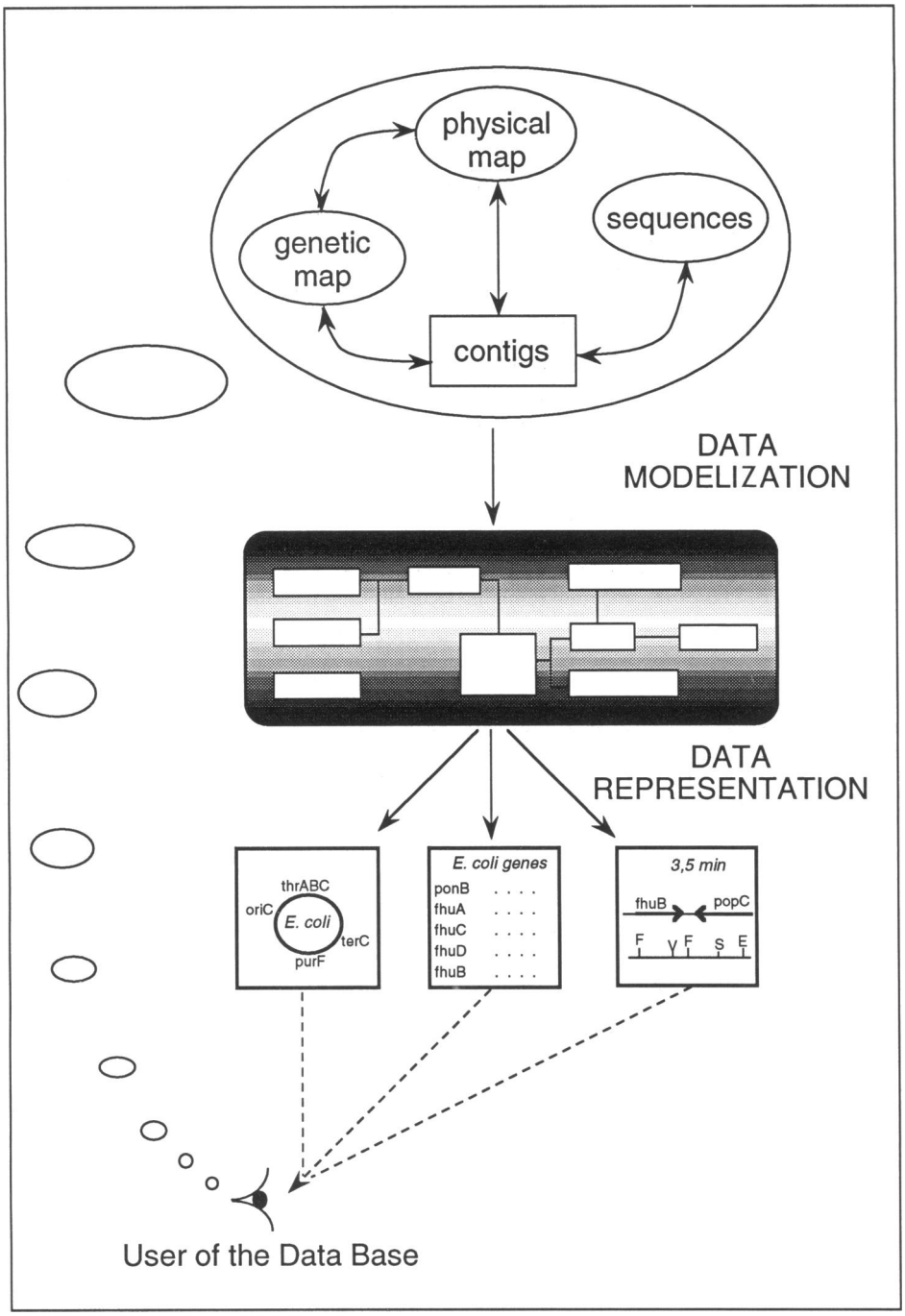

FIG. 3. User's views of the Colibri data base. The logical structure of the data in a data base must express, in the best possible way, the biologist's mental representation of these data (data modelization step). To make this organization quite invisible to the user, one defines a set of layouts presenting data on the screen (data representation step).

4B). The direction of gene transcription is indicated by the direction of the arrows. Names of the gene corresponding to the coding regions are localized in sensitive areas: by clicking twice with the mouse on one of them, the user can see all the information of the corresponding gene. In this graphical representation, numbers associated with each gene were obtained after a statistical analysis intended to define *E. coli* classes according to their codon usage. What, precisely, are these gene classes?

Figure 5 is a graphic representation of the codon usage of all *E. coli* genes. It was obtained by a method called factorial correspondence analysis (12). In this graph, a point is a gene and two points appear as neighbors if the corresponding genes have a similar codon usage. Using a second method that automatically clusters the CDSs that are close to one another (8), we have identified three well-separated classes (Table 1). As seen in Fig. 5, the three classes of *E. coli* genes can be distinguished by their biological properties. The two

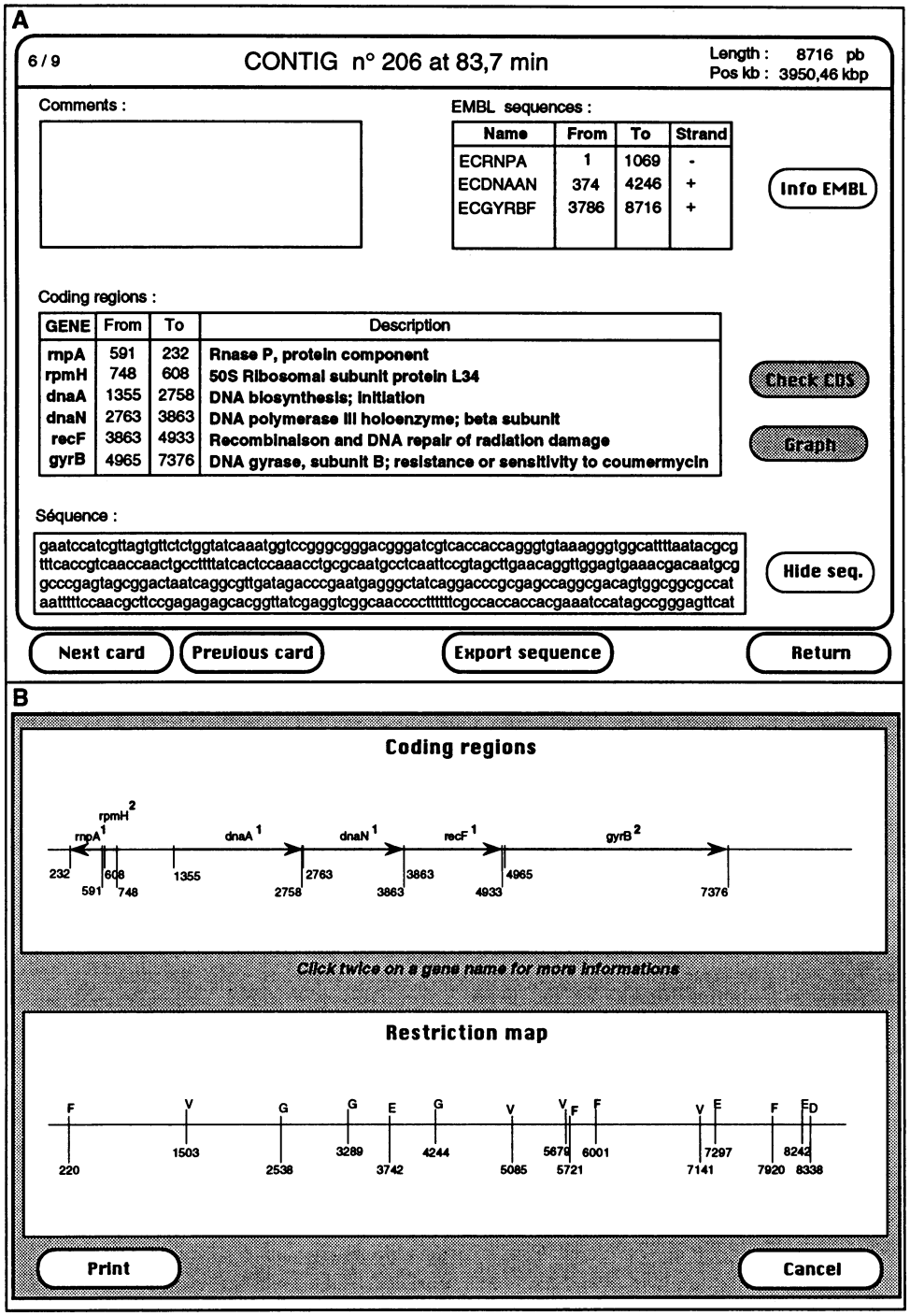

FIG. 4. Consultation of data for a recorded contig localized at 83.7 min. (A) Layout of the 4D file [CONTIG] presenting data (or fields) from the file [CONTIG] but also from the files [Sequences], [Coding R], and [GENES] (see Fig. 2). (B) By clicking on the "Graph" button, the user can see a graphical representation of the position of the coding regions and the restriction enzymes sites on the contig (see text).

first classes, C1 and C2, have already been identified from a limited set of genes (3, 9); the new finding is that to describe properly the codon usage of all *E. coli* genes, it has been necessary to introduce a third class (C3; Fig. 5). Genes in this third class code for fimbriae, flagella and pili, and integration host factors; they also comprise genes controlling cell division. In addition, the third class contains genes that encode insertion sequences. A few genes, such as the *mut* genes, are found at the border between classes and may therefore change class from one release of Colibri to the next one. It has also been found that the codon bias, on the one hand, and di- to pentaoligonucleotide bias, on the other hand, are specific to these three classes (21).

Results of this statistical analysis are then recorded in the data base (see [Codon usage] 4D file; Fig. 2). Thus, we show in Fig. 6 the *E. coli* gene information in a list form. This

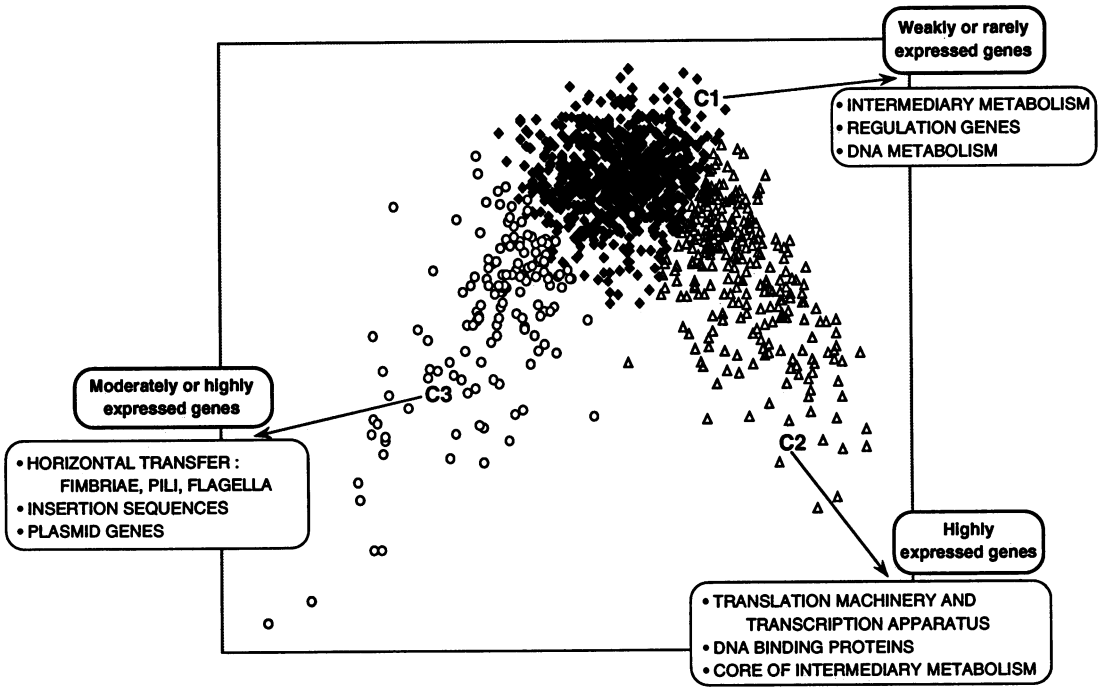

FIG. 5. Factorial correspondence analysis of *E. coli* genes for a clustering into three classes. Each CDS is represented as a point in a 61-dimensional space, each dimension corresponding to the relative frequency of one of the 61 codons. The set of CDSs appears as a cloud of points in the space of codon frequencies. Genes that have a similar codon usage will therefore appear as neighbors. Clustering is then performed by a second method that automatically clusters objects that are close to one another. Here, genes are represented by a solid diamond, an open triangle, or an open circle according to the class to which they belong (classes 1, 2, and 3, respectively).

selection was obtained by searching genes linked to the keyword "DNA biosynthesis." As shown in the Class column, these genes belong to the C1 class from a codon usage point of view, apart from the *dnaKJ* operon. Other information shown in this layout includes the name, the experimental genetic map position (noted as M_G for Genetic Minutes), and the map position in minutes calculated from the genomic address (column Pos_kb) of the gene transcription on the *E. coli* chromosome (noted as M_P for Physical Minutes, with M_P = kb * 100/ 4,719.6). These physical minutes allow adjustment of the experimental ones defined by Bachmann (1). Moreover, we indicate the direction of gene transcription with respect to the replication forks (OriC column). From this selected gene records, the user has access to the corresponding EMBL entries and bibliographic references (>>EMBL button), to the corresponding contigs (>>CONTIG button), or to the results of the FASTP routines (30), if any (>>Scan FastP button; see below).

Calling External Procedures

One of the main goal of scientists involved in sequencing has been to identify gene products (proteins, RNA, DNA regulatory sequences, etc.) in order to associate them formally with their genetic and physiological properties. The most common operation is therefore identification of a coding sequence, automatic translation into a polypeptide sequence according to a given genetic code, and comparison of the latter with known sequences present in data banks. Computers provide great assistance in identifying open reading frames, translating sequences, finding similarities with data banks, aligning sequences, etc. Therefore we have developed on the Macintosh computer several software analyses, generally written in C, permitting such standard treatment of sequences. Most of these algorithms could have been written in 4D language, but their slow running makes them unusable in practice. For this reason, integration of external procedures in Colibri has been used (22) for sequence analyses such as identification of coding regions, translation of CDSs, and establishment of restriction maps with the eight restriction enzymes used by Kohara et al. (16): in such cases, the user simply has to select the sequence(s) to be treated in the data base. If no particular choice motivated by biological expertise is required, results are then automatically recorded in appropriate fields of the data base. A more complex routine used for rapid similarity searching in the proteins data bank (FastP) (30) has also been integrated. In that case, parameters necessary to run the program are numerous. Therefore, using the 4D system facilities, we have developed an interface allowing one to select from Colibri the query sequence, the protein data bank to be scanned, a distance matrix between amino acids, and finally the number of data bank sequences to be kept for further study. Results of the scanning are automatically loaded in the data base: for each selected protein of the data bank, its similarity score with the query sequence and information such as its description, its keywords, and its bibliographic references are recorded. The scores obtained with all the data bank entries are finally recorded to draw the corresponding histogram (the user generally wants to see the position of the selected sequences in comparison with the whole results of the FastP routine). Organization of these data in our data base (Fig. 2, [FastP] 4D file) allows the user to readily consult the results obtained from each running of this external routine.

TABLE 1. Genetic map, physical map, gene transcription orientation, and codon usage class

Name	Length (bp)	Map position		oriC[c]	Class[d]	Description
		M_G[a]	M_P[b]			
aceA	1,305	91	91.05	d	2	Acetate; utilization of acetate; isocitrate lyase (EC 4.1.3.1)
aceB	1,602	91	91.01	d	1	Acetate; utilization of acetate; malate synthase A (EC 4.1.3.2)
aceE	2,658	3	2.65	d	2	Acetate; acetate requirement; pyruvate dehydrogenase (decarboxylase component)
aceF	1,890	3	2.71	d	2	Acetate; acetate requirement; pyruvate dehydrogenase (dihydrolipoyltransacetylase component)
aceK	1,737	91	91.08	d	1	isocitrate dehydrogenase kinase/phosphatase
ackA	1,203	50	51.37	d	2	Acetate kinase activity (EC 2.7.2.1)
acn	2,676	28	28.65	d	1	Aconitate hydratase
acpP	237	24	24.78	d	2	Acyl carrier protein
ada	1,065	48	49.13	d	1	Inductible DNA repair system protecting against methylating and alkylating agents; O^6-methylguanine-DNA-methyltransferase
add	999	36	36.43	i	1	Adenosine deaminase (EC 3.5.4.4)
adhE	2,676	27	27.76	i	2	CoA[e]-linked acetaldehyde dehydrogenase and alcohol dehydrogenase
adk	645	11	10.79	d	2	Phospholipid synthesis; adenylate kinase activity (EC 2.7.4.3); pleiotropic effects on glycerol-3-phosphate; acetyltransferase activity
agp	1,242		22.89	d	1	Periplasmic acid glucose-1-phosphatase
alaS	2,628	58	60.01	d	2	Alanyl-tRNA synthetase (EC 6.1.1.7)
ald	1,440	31	31.92	d	1	Lactaldehyde dehydrogenase
aldH	1,488	31	29.22	d	1	Aldehyde dehydrogenase, NAD linked
alkA	849	45	45.03	nc	3	3-Methyladenine DNA glycosylase II, inducible
alkB	651	47	49.12	d	1	DNA repair system specific for alkylated DNA
amn	1,452	43	43.38	i	1	AMP nucleosidase (EC 3.2.2.4)
ampC	1,134	94	94.46	i	1	β-Lactamase; penicillin resistance
ampD	549		2.62	i	1	β-Lactamase regulation; cytoplasmic protein
ampE	852		2.60	i	1	β-Lactamase regulation; inner membrane protein
ams	2,448	24	24.46	nc	2	RNase; RNase E activity; alteration of mRNA stability
anr	339		39.00	d	1	Activator of *ntr*-like gene
ansA	969	39	39.62	i	1	L-Asparaginase I, cytoplasmic
ansB	1,047		65.98	d	2	L-Asparaginase II
ant	1,086	0	0.39	d	1	Na^+/H antiporter activity
apaG	375		1.09	i	1	Function unknown
apaH	840	1	1.07	i	1	Diadenosine tetraphosphatase
appA	1,299	22	22.71	nc	1	pH 2.5 acid phosphatase; exopolyphosphatase (EC 3.6.1.11)
appY	729		12.63	d	3	Transcriptional regulatory protein
apt	552	11	10.66	d	2	Adenine phosphorobosyltransferase (EC 2.4.2.7)
araA	1,701	1	1.45	i	1	Arabinose; L-arabinose isomerase (EC 5.3.1.4)
araB	1,503	1	1.42	i	1	Arabinose; ribulokinase (EC 2.7.1.16)
araC	879	1	1.49	d	1	Arabinose; regulatory gene; activator and repressor protein
araD	696	1	1.40	i	1	Arabinose; L-ribulosephosphate 4-epimerase (EC 5.1.3.4)
araE	1,419	61	63.45	d	1	Arabinose; low-affinity L-arabinose transport system; L-arabinose proton symport
araF	987	45	45.30	i	1	Arabinose; L-arabinose binding protein
araG	1,512	45	45.32	i	1	Arabinose; high-affinity L-arabinose transport system
araH	987	45	45.35	i	1	Arabinose; high-affinity L-arabinose transport system; membrane protein
araJ	1,182	9	8.99	i	1	Arabinose
arcB	2,337		72.42	d	1	Acridine; involved in the regulation of F pilus synthesis and enzymes involved in aerobic metabolism
argA	1,329	61	62.79	i	1	Arginine; amino acid acetyltransferase; *N*-acetylglucosamine synthase (EC 2.3.1.1)
argB	777	90	89.88	d	1	Arginine; acetylglutamate kinase (EC 2.7.2.8)
argC	1,005	90	89.85	d	1	Arginine; *N*-acetyl-gamma-glutamyl-phosphate reductase (EC 1.2.1.38)
argD	1,221	74	75.43	i	1	Arginine; acetylomithine δ-aminotransferase (EC 2.6.1.11)
argE	1,152	90	89.83	i	1	Arginine; acetylornithine deacetylase (EC 3.5.1.16)
argF	1,002	7	6.40	i	1	Arginine; ornithine carbamoyltransferase (duplicate gene) (EC 2.1.3.3)
argG	1,344	69	71.74	i	2	Arginine; argininosuccinate synthetase (EC 6.3.4.5)
argH		90	89.90	ns		Arginine; argininosuccinate lyase (EC 4.3.2.1)
argI	1,002	97	96.48	d	1	Arginine; ornithine carbamoyltransferase (duplicate gene) (EC 2.1.3.3)
argR	468	71	73.15	nc	1	Arginine; repressor of *arg* regulon
argS	1,731	40	41.75	i	2	Arginine; arginyl-tRNA synthetase (EC 6.1.1.19)
argT		50	51.72	ns		Arginine; sequence homologous to *argT* of *S. typhimurium*, which codes for lysine-, arginine-, ornithine-binding protein
aroA	1,281	20	20.58	nc	1	Aromatic; 3-enol-pyruvylshikimate-5-phosphate synthase (EC 2.5.1.19)
aroB	1,086	75	76.04	d	1	Aromatic; dehydroquinate synthase (EC 4.6.1.3)
aroC	1,071	51	52.13	d	1	Aromatic; chorismate synthase (EC 4.6.1.4)
aroD	720	37	37.98	i	1	Aromatic; 5-dehydroquinate dehydratase (EC 4.2.1.10)
aroE	816	72	74.34	nc	3	Aromatic; dehydroshikimate reductase (EC 1.1.1.25)

Continued on following page

TABLE 1—*Continued*

Name	Length (bp)	Map position		*oriC*[c]	Class[d]	Description
		M.G[a]	M.P[b]			
aroF	1,068	57	58.22	nc	1	Aromatic; DAHP synthetase (EC 4.1.2.15) (tyrosine repressible)
aroG	1,053	17	16.89	nc	1	Aromatic; DAHP synthetase (EC 4.1.2.15) (phenylalanine repressible)
aroH	1,047	37	38.11	i	1	Aromatic; 3-deoxy-D-arabino heptulosonate 7-phosphate (DAHP) synthetase (EC 4.1.2.15) (tryptophan repressible)
aroK	435	75	76.07	d	1	Aromatic; shikimic acid kinase I
aroL	525	9	8.87	d	3	Aromatic; shikimate kinase II (EC 2.7.1.71)
aroM	678	9	8.89	d	1	Aromatic; unknown function; regulated by *aroR*
aroP	1,368	3	2.57	d	1	Aromatic; general aromatic amino acid transport
ascB	1,425		60.47	i	1	Phospho-β-glucosidase
ascF	1,458		60.44	i	1	PTS enzyme II-*asc*
ascG	1,005		60.41	d	1	*asc* repressor
asdX	1,101	76	77.27	d	1	Aspartate-semialdehyde dehydrogenase (EC 1.2.1.11)
aslA	1,428			nm	1	Arylsulfatase
aslB	1,236			nm	1	Arylsulfatase regulator
asnA	993	84	84.88	d	1	Asparagine; asparagine synthetase A (EC 6.3.1.1)
asnB	1,665	16	15.16	i	2	Asparagine; asparagine synthetase B (EC 6.3.1.1)
asnS	1,401	21	21.22	i	2	Asparagine; asparaginyl-tRNA synthetase (EC 6.1.1.22)
aspA	1,434	94	94.23	i	2	Aspartate; L-aspartate ammonia-lyase (aspartase) (EC 4.3.1.1)
aspC	1,188	21	21.14	nc	1	Aspartate; aspartate aminotranferase (EC 2.6.1.1)
aspS	1,773		41.54	d	2	Aspartate; aspartyl-tRNA synthetase
avtA	678	84	80.78	d	3	Alanine-alpha-ketoisovalerate transaminase, transaminase C
barA	2,757			nm	1	Sensor regulator protein
BCCP	471	71	73.60	i	2	Fatty acid biosynthesis; subunit of acetyl-CoA carboxylase; biotin carboxyl carrier protein
bcp	471		55.13	i	1	Bacterioferritin comigratory protein
betA	1,668	7	7.14	i	1	Betaine; choline dehydrogenase
betB	1,470	7	7.18	i	1	Betaine; betaine aldehyde dehydrogenase
betI	585	7	7.21	i	1	Betaine; function unknown
betT	2,031	7	7.22	d	1	Betaine; high-affinity choline transport
bfr	477		74.97	d	1	Bacterioferritin
bglB	1,416	84	84.36	d	1	β-Glucoside; phospho-β-glucosidase B
bglC	837	84	84.43	d	3	β-Glucoside; β-glucoside transport
bglS	1,878	84	84.39	d	1	β-Glucoside; β-glucoside transport; positive regulatory gene
bioA	1,293	17	17.41	i	1	Biotin; 7,8-diamino-pelargonic acid aminotransferase (synthetase for Bachmann [1])
bioB	1,041	17	17.44	d	1	Biotin; biotin synthetase; conversion of dethiobiotin to biotin
bioC	756	17	17.49	d	1	Biotin; block prior to pimeloyl-CoA
bioD	660	17	17.50	d	3	Biotin; dethiobiotin synthetase
bioF	1,155	17	17.46	d	1	Biotin; 7-keto-8-amino pelargonic acid synthetase
bioH	768	75	76.61	d	1	Biotin; block prior to pimeloyl-CoA
birA	966	90	90.10	d	1	Biotin retention; biotin-[acetyl-CoA carboxylase] holoenzyme synthetase; biotin operon repressor
bisC	2,181	80	80.22	d	1	Biotin sulfoxide; biotin sulfoxide reductase, structural gene
bolA	348		9.87	d	3	Function unknown
btuB	1,845	90	89.90	d	1	B12 uptake; receptor for vitamin B_{12}, E colicins, and bacteriophage BF23
btuC	879	37	38.37	d	3	B12 uptake; vitamine B_{12} transport
btuD	750	37	38.34	d	3	B12 uptake; vitamin B_{12} transport mechanism; peripheral membrane component
btuE	552	37	38.36	d	1	B12 uptake; vitamin B_{12} transport mechanism; possible periplasmic protein
btuR	591	28	28.45	nc	1	B12 uptake; vitamin B_{12} transport mechanism; regulatory gene affecting *btuB*; outer membrane protein
cadA	2,145	94	94.01	i	2	Cadaverine; lysine decarboxylase (EC 4.1.1.18)
cadB	1,332		94.05	i	2	Cadaverine
cadC	1,536		94.09	i	3	Cadaverine; required for Pcad induction; transcriptional activator
capR	2,352	10	9.98	d	1	Long form; DBA-binding, ATP-dependent protease La
carA	1,149	1	0.67	d	1	Pyrimidine; carbamoyl-phosphate synthase (EC 2.7.2.9), glutamine (light) subunit
carB	3,222	1	0.69	d	2	Pyrimidine; carbamoyl-phosphate synthase (EC 2.7.2.9); ammonia (heavy) subunit
cca	1,239	67	69.02	i	1	tRNA nucleotidyl transferase (EC 2.7.7.25)
cdd	948	46	47.21	nc	1	Deoxycytidine deaminase (EC 3.5.4.5)
cdh	750	89	88.77	d	3	CDP-diglyceride hydrolase
cds	750	4	4.40	d	1	CDP-diglyceride synthetase (CTP: phosphatidate cytidylyltransferase) (EC 2.7.7.4)
celA	318		38.99	d	1	Cellobiose
celB	1,251	38	38.96	d	1	Cellobiose; phosphoenolpyruvate-dependent phosphotransferase enzyme II-cellobiose; transport of cellobiose, arbutin, and salicin
celC	348	38	38.95	d	1	Cellobiose; phosphoenolpyruvate-dependent phosphotransferase enzyme III-cellobiose; transport of cellobiose, arbutin, and salicin
celD	840	38	38.93	d	1	Cellobiose; negative regulatory gene of the *cel* operon
celF	1,116	38	38.90	d	1	Cellobiose; phospho-β-glucosidase
cet	1,347	100	99.92	d	1	Colicin E2; tolerance to colicin E2

Continued on following page

TABLE 1—*Continued*

Name	Length (bp)	Map position		oriC[c]	Class[d]	Description
		M_G[a]	M_P[b]			
cheA	1,965	42	42.05	d	1	Chemotaxis; chemotactic response
cheB	1,050	41	41.91	d	1	Chemotaxis; chemotactic response; methylesterase activity
cheR	861	41	41.93	d	1	Chemotaxis; chemotactic response; methylesterase activity
cheW	504	42	42.04	d	1	Chemotaxis; chemotactic response
cheY	390	41	41.90	d	1	Chemotaxis; chemotactic response
cheZ	645	41	41.89	d	1	Chemotaxis; chemotactic response
chlD	903	17	17.15	nc	1	Chlorate; molybdenum uptake; nitrate reductase, formate dehydrogenase, and biotin sulfoxide reductase activity
chlE	1,236	18	18.62	i	1	Chlorate; molybdopterin biosynthesis; nitrate reductase, formate dehydrogenase, and biotin sulfoxide reductase activity; regulatory function
chlN	750	18	18.60	i	1	Chlorate; molybdopterin biosynthesis; nitrate reductase, formate dehydrogenase, and biotin sulfoxide reductase activity
cir	1,992	43	47.78	d	1	Production of colicin I receptor
clpA	2,277		19.85	d	1	ATP-dependent clp protease
clpB	2,571		57.97	d	1	ATP-dependent protease binding subunit
clpG	837			nm	3	Surface antigen
clpP	624		9.97	nc	1	ATP-dependent *clp* protease proteolytic component
cmlA	642	19	19.96	nc	3	Chloramphenicol; chloramphenicol acetyltransferase; resistance or sensitivity to chloramphenicol
coaA	1,011	90	90.13	nc	3	Pantothenate kinase (EC 2.7.1.33); uncharacterized growth defect
codA	1,284	8	7.80	nc	1	Cytosine deaminase (EC 3.5.4.1)
codB	1,260	8	7.77	nc	1	Cytosine permease; cytosine transport
colA	444		47.84	nc	2	Colicin A lysis protein
cpdB	1,941	96	95.67	i	1	2′,3′-Cyclic-nucleotide 2′-phosphodiesterase (EC 3.1.4.16)
cpxA	1,374	89	88.64	nc	1	F-pilus formation, surface exclusion, conjugal donor activity
crp	633	74	75.37	i	2	Cyclic AMP receptor protein
crr	510	52	53.94	d	2	Phosphocarrier protein for glucose of the PTS system; glucose phosphotransferase system enzyme IIIglc, structural gene
cscB	1,248			nm	3	Sucrose permease
cspA	213		79.41	nc	2	Cold shock protein
cstA	1,686		13.69	d	1	Carbon starvation gene product
cutE	1,539		14.96	i	1	Involved in copper homeostasis
cya	2,544	86	86.23	d	1	Adenylate cyclase (EC 4.6.1.1)
cybB	525	17	31.97	nc	3	Cytochrome b_{561}
cydI	1,572	17	16.65	d	2	Cytochrome oxydase d subunit I precursor
cydII	1,140	17	16.69	d	2	Cytochrome oxydase d subunit II
cynR	900		7.83	i	1	Cyanate; positive regulatory protein for the cyn operon
cynS	471	8	7.85	d	1	Cyanate; cyanate aminohydrolase (EC 3.5.5.3); cyanase
cynT	657	8	7.84	d	1	Cyanate; cyanate permease
cynX	1,053	8	7.86	d	1	Cyanate; function unknown
cyoA	948	10	9.80	i	2	Cytochrome *o* terminal oxidase complex
cyoB	1,992	10	9.76	i	2	Cytochrome *o* terminal oxidase complex
cyoC	615	10	9.75	i	2	Cytochrome *o* terminal oxidase complex
cyoD	330	10	9.74	i	2	Cytochrome *o* terminal oxidase complex
cyoE	891	10	9.72	i	1	Cytochrome *o* terminal oxidase complex
cysA	1,098	52	54.09	nc	1	Cysteine; sulfate permease; chromate resistance
cysB	972	28	28.60	d	1	Cysteine; positive regulatory gene for cycteine biosynthesis
cysC	603	59	61.20	d	1	Cysteine; adenosine 5′-phosphosulfate kinase (EC 2.7.1.25)
cysD	906	59	61.24	d	1	Cysteine; ATP sulfurylase (ATP:sulfate adenyltransferase) (EC 2.7.7.4)
cysE	822	81	81.64	nc	1	Cysteine; serine acetyltransferase (EC 2.3.1.30)
cysG	1,371	74	75.86	nc	1	Cysteine; sulfite reduction and possible nitrite reduction; siroheme synthesis
cysH	735	59	61.46	nc	1	Cysteine; adenylylsulfate reductase (EC 1.8.99.2)
cysI	1,719	59	61.42	nc	1	Cysteine; NADPH-sulfite reductase (EC 1.8.1.2), alpha subunit
cysJ	1,800	59	61.38	nc	1	Cysteine; NADPH-sulfite reductase (EC 1.8.1.2), beta subunit
cysK	972	52	54.00	d	1	Cysteine; cysteine synthetase; *O*-acetylserine sulfhydrolase A (EC 4.2.99.8)
cysM	912	52	54.12	nc	1	Cysteine; *O*-acetylserine (thiol)-lyase-B; *O*-acetylserine sulfhydrolase B (EC 4.2.99.8)
cysN	1,425	59	61.21	d	1	Cysteine; ATP-sulfurylase (ATP:sulfate adenylyltransferase) (EC 2.7.7.4), subunit
cysP	1,017		54.04	nc	2	Cysteine; thiosulfate-binding protein
cysQ	741		95.71	d	1	Cysteine; ammonium transport protein
cysS	1,386	12	12.02	d	1	Cysteine; cysteinyl-tRNA synthetase (EC 6.1.1.16)
cysT	834	42	54.06	nc	1	Cysteine; cysteine tRNA; sulfate permease
cysW	876		54.08	nc	1	Cysteine; sulfate permease
cysX	393		81.64	nc	1	Cysteine; function unknown
cysZ		52	54.04	ns		Cysteine; *O*-acetylserine (thiol)-lyase-A
cytR	1,023	89	89.06	i	1	Regulatory gene for *deo* operon, *udp*, and *cdd*
dacA	1,209	15	14.39	i	2	D-Alanine carboxypeptidase, fraction A; penicillin-binding protein 5

Continued on following page

TABLE 1—*Continued*

Name	Length (bp)	Map position		*oriC*[c]	Class[d]	Description
		M_G[a]	M_P[b]			
dacB	1,434	69	71.95	i	1	D-Alanine carboxypeptidase, fraction B; penicillin-binding protein 4 (PBP4); involved as a DD-carboxypeptidase endopeptidase in murein metabolism
dacC	1,200		18.96	d	1	D-Alanine carboxypeptidase, fraction C; penicillin-binding protein 6
damX	834	74	75.99	d	1	DNA; DNA adenine methylase which methylates the sequence GATC
dapA	879	53	55.07	d	1	Diaminopimelate; dihydrodipicolinate synthase (EC 4.2.1.52)
dapB	822	1	0.64	d	1	Diaminopimelate; dihydrodipicolinate reductase (EC 1.3.1.26)
dapD	825	4	4.10	nc	2	Diaminopimelate; tetrahydrodipicolinate *N*-succinyltransferase
dapE	1,128	53	54.94	i	1	Diaminopimelate; *N*-succinyl-diaminopimelate deacylase
dapF	825	86	86.30	d	1	Diaminopimelate; diaminopimelate epimerase
dbpA	1,296		30.22	nc	1	DEAD box protein
dcd	582	46	46.44	nc	1	2′-Deoxycytidine 5′-triphosphate deaminase activity (EC 3.5.4.13)
dcm	1,416	43	43.27	d	1	DNA cytosine methylase
dcp	2,046	29	28.91	d	1	Dipeptidyl carboxypeptidase
ddl	921	2	2.22	d	1	D-alanine: D-alanine ligase; mureine
ddlA	1,095		8.60	nc	1	D-alanine:D-alanine ligase A
deaD	1,716		71.46	d	2	Encodes a presumed ATP-dependent RNA helicase
dedA	660		51.86	d	1	Folate; function unknown
dedB	915		51.84	d	1	Folate; folate coenzyme
dedC	1,272	50	51.81	d	1	Folate; folylpolyglutamate-dihydrofolate synthetase
dedD	636		51.79	d	1	Folate; function unknown
dedE	489		51.78	d	1	Folate; function unknown
dedF	570		51.73	d	1	Folate; function unknown
deoA		100	99.50	ns		Deoxyribose; thymidine phosphorylase (EC 2.4.2.4)
deoB		100	99.46	ns		Deoxyribose; phosphopentomutase (EC 2.7.5.6)
deoC	777	100	99.48	nc	2	Deoxyribose; deoxyribose 5-phosphate aldolase (EC 4.1.2.4)
deoD	720	100	99.46	nc	2	Deoxyribose; purine-nucleoside phosphorylase (EC 2.4.2.1)
deoR	756	19	18.99	i	1	Deoxyribose; regulatory gene for *deo* operon
dgk	369	92	91.87	d	1	Diglyceride kinase
dgt	1,518		3.85	d	1	dGTP triphosphohydrolase
dicA	405	35	35.29	i	3	DNA-binding protein; regulatory gene
dicB	330	35	35.32	i	3	Control cell division; DNA-binding protein
dicC	228	35	35.29	d	3	DNA binding protein; regulatory gene
dinG	1,914		17.95	d	1	Function unknown
div			51.94	ns		Function unknown
divE	657	22	22.15	i	1	Division; membrane protein biosynthesis
dksA	453		3.47	d	1	*dnaK* suppressor
dld	1,713	47	47.26	i	1	D-Lactate dehydrogenase (EC 1.1.1.28)
dmsA	2,358	20	20.22	d	1	Anaerobic dimethyl sulfoxide reductase, subunit A
dmsB	624	20	20.27	d	1	Anaerobic dimethyl sulfoxide reductase, subunit B
dmsC	864	20	20.29	d	1	Anaerobic dimethyl sulfoxide reductase, subunit C
dnaA	1,404	83	83.82	d	1	DNA; DNA biosynthesis; initiation
dnaB	1,416	92	91.99	nc	1	DNA; DNA biosynthesis; chain elongation
dnaC	738	99	99.15	i	1	DNA; DNA biosynthesis; initiation and chain elongation
dnaE	3,483	4	4.60	d	1	DNA; DNA polymerase III alpha subunit
dnaG	1,743	67	69.22	i	1	DNA; DNA biosynthesis; primase
dnaJ	1,131	0	0.32	d	2	DNA; DNA biosynthesis
dnaK	1,917	0	0.27	d	2	DNA; DNA biosynthesis; heat shock protein
dnaN	1,101	83	83.80	d	1	DNA; DNA biosynthesis; DNA polymerase III holoenzyme beta subunit
dnaT	540	99	99.17	i	1	DNA; DNA biosynthesis; primasomal protein i
dnaZX	1,932	11	10.68	d	1	DNA; DNA biosynthesis; DNA polymerase III gamma subunit; DNA elongation factor III
dniR	669		5.19	i	1	Involved in hexaheme nitrite reductase expression
dpj	381		57.29	d	1	Function unknown
dsdA		51	52.63	ns		D-Serine; D-serine deaminase
dsdC	795	51	52.65	nc	3	D-Serine; regulatory gene for *dsdA* (activator)
dut	453	82	82.34	i	2	dUTPase; deoxyuridine triphosphatase (EC 3.6.1.33)
dye	717	100	99.95	i	2	Negative regulatory gene of genes in aerobic pathways
eae	2,820			nm	3	Attaching and effacing gene
ebgA	3,108	67	69.48	i	1	Phospho-β-D-galactosidase, alpha subunit; cryptic gene
ebgB		67	69.56	ns		Possible homolog of *lacY*; in *ebg* operon
ebgC	516	67	69.55	i	1	Phospho-β-D-galactosidase, beta subunit; cryptic gene
ebgR	984	67	69.46	i	1	Regulatory gene of *ebg* operon; repressor protein
eda	642	41	41.21	nc	2	2-Keto-3-deoxygluconate 6-phosphate aldolase (EC 4.1.2.14)
edd	1,809	41	41.17	nc	1	6-Phosphogluconate dehydratase (EC 4.2.1.12)
emrA1	1,173		59.91	nc	1	Multidrug resistance
emrA2	1,539		59.94	nc	1	Multidrug resistance
endA	708	64	65.78	i	1	DNA; DNA-specific endonuclease I

Continued on following page

TABLE 1—*Continued*

Name	Length (bp)	Map position		*oriC*[c]	Class[d]	Description
		M_G[a]	M_P[b]			
eno		60	61.93	ns		Enolase (EC 4.2.1.11)
entB		14	13.59	ns		Enterochelin; 2,3-dihydro-2,3-dihydroxybenzoate synthetase
entC	1,173	14	13.53	nc	1	Enterochelin; isochorismate synthetase
entD	767	13	13.19	d		Enterochelin; enterochelin synthetase, component D
entE	1,608	14	13.56	nc	1	Enterochelin; enterochelin synthetase, component E
entF	3,882	14	13.33	d	1	Enterochelin; enterochelin synthetase, component F
envA	918	2	2.31	d	1	Envelope; cell envelope and cell separation
envC	1,155	81	81.62	nc	1	Envelope; envelope protein; anomalous cell division; chain formation
envD	2,895	81	81.65	nc	1	Envelope; cell division
envY	759	13	12.71	nc	1	Envelope; envelope protein; thermoregulation of porin synthesis
envZ	1,188	75	76.42	d	1	Envelope; production of outer membrane proteins; regulatory gene
era	951		57.33	d	2	RAS-like protein
ermBC	738			nm	3	23S rRNA methylase
exbB	735	65	67.92	d	1	Uptake of enterochelin; resistance or sensitivity to colicins
exbD	426	65	67.91	d	1	Uptake of enterochelin; resistance or sensitivity to colicins
fabA	516	22	21.84	d	2	Fatty acid biosynthesis; β-hydroxydecanoyl thioester dehydratase (EC 4.2.1.60)
fabB	1,221	50	52.00	d	2	Fatty acid biosynthesis; β-ketoayl-[acyl-carrier-protein] synthase I (EC 2.3.1.41)
fabD	930	24	24.74	d	1	Fatty acid biosynthesis; malonyl-CoA-[acyl-carrier protein] transacylase (EC 2.3.1.39)
fabE	1,350	71	73.61	i	2	Fatty acid biosynthesis; subunit of acetyl-CoA carboxylase (EC 6.4.1.2); biotin carboxylase
fabG	735	24	24.76	d	2	Fatty acid biosynthesis; 3-ketoacyl-acyl carrier protein reductase; biotin carboxylase
fabH	954		24.72	d	1	Fatty acid biosynthesis; β-ketoacyl-acyl carrier protein synthase III
fadA	1,161	87	87.02	i	1	Fatty acid degradation; 3-ketoacyl-CoA thiolase I (EC 2.3.1.16)
fadB	2,187	87	87.04	i	1	Fatty acid degradation; L-3-hydroxyacyl-CoA dehydrogenase (EC 1.1.1.35); δ-3-*cis*-δ-2-*trans*-enoyl-CoA isomerase (EC 5.3.3.8); 3-hydroxyacyl-CoA epimerase (EC 5.1.2.3); enoyl-CoA and enoyl-CoA hydratase (EC 4.2.1.17)
fadI	702		86.99	d	1	Fatty acid degradation; regulatory gene: activator of *fadB* and *fadE* genes; flavin oxidoreductase
fadL	1,347	51	52.43	i	1	Fatty acid degradation; transport of long-chain fatty acids and sensitivity to phage T2
fadR	717	26	26.45	nc	1	Fatty acid degradation; negative regulatory gene for *fad* regulon and *aceBA* operon
fda	1,077	63	65.34	d	2	Fructose 1,6-bisphosphate aldolase (EC 4.1.2.13)
fdhE	933		88.50	nc	1	Soluble protein involved in formation of formate dehydrogenase [FDH(N)]
fdhF	1,680	93	92.75	i	2	Formate dehydrogenase (formate hydrogen-lyase linked), selenopolypeptide
fdnG	585		33.15	i	1	Nitrate-inductible formate dehydrogenase
fdnH	2,412		33.16	i	2	Nitrate-inductible formate dehydrogenase
fdnI	882		33.21	i	1	Nitrate-inductible formate dehydrogenase
fdp	996	96	96.10	d	2	Fructose-1,6-bisphosphatase (EC 3.1.3.11)
fdx	336			nm	2	Ferredoxin (2FE-2S) protein
fecA	2,325	93	97.33	i	1	Iron; citrate-dependent iron transport, outer membrane receptor
fecB	903	8	97.28	i	1	Iron; citrate-dependent iron transport, periplasmic protein
fecC	999		97.26	i	1	Iron; citrate-dependent iron transport
fecD	957	8	97.24	i	1	Iron; citrate-dependent iron transport, membrane-bound protein
fecE	768		97.22	i	3	Iron; function unknown
fecI	522		97.36	i	1	Iron; function unknown
fecR	954		97.35	i	1	Iron; function unknown
fepA	2,235	13	13.24	i	1	Iron; receptor for ferrienterochelin and colicins B and D; enterochelin-dependent iron transport
fepB	957	13	13.52	nc	1	Iron; ferric enterobactin (enterochelin) uptake; periplasmic component
fepC	813	13	13.47	nc	1	Iron; ferric enterobactin transport protein, ATP-binding protein; cytoplasmic membrane component
fepD	1,002	13	13.43	nc	1	Iron; ferric enterobactin transport protein
fepG	990	13	13.45	nc	1	Iron; ferric enterobactin transport protein
fes	1,125	13	13.30	d	1	Iron; enterochelin esterase
fhlA	2,061	58	60.77	i	1	Transcriptional activator of the formate hydrogen-lyase; possible electron transport system
fhuA	2,244	4	3.59	d	2	Ferric hydroxamate uptake and T1; outer membrane protein receptor for ferrichrome, colicin M and phages T1, T5, and ϕ80
fhuB	1,980	4	3.68	d	1	Ferric hydroxamate uptake; hydroxamate-dependent iron uptake, cytoplasmic membrane component
fhuC	798	4	3.64	d	1	Ferric hydroxamate uptake; hydroxamate-dependent iron uptake, cytoplasmic membrane component; involved in Fe^{3+} transport
fhuD	891	4	3.66	d	1	Ferric hydroxamate uptake; hydroxamate-dependent iron uptake, cytoplasmic membrane component; involved in $Fe3^{+}$ transport

Continued on following page

TABLE 1—*Continued*

Name	Length (bp)	Map position		*oriC*[c]	Class[d]	Description
		M_G[a]	M_P[b]			
fhuE	2,187	16	24.95	i	1	Ferric hydroxamate uptake; outer membrane receptor for ferric-rhodotorulic acid
fic	600	74	75.48	d	1	Filamentation in presence of cyclic AMP in mutant
fimA	543	98	97.99	nc	1	Fimbriae; type 1 fimbrin (pilin), structural gene
fimB	600	98	97.94	nc	3	Fimbriae; regulatory gene for expression of *himA*
fimC		98	98.02	ns		Fimbriae; biosynthesis of type 1 fimbriae
fimD	2,631	98	98.02	d	3	Fimbriae; biosynthesis of type 1 fimbriae
fimE	594	98	97.97	nc	3	Fimbriae; regulatory gene for expression of *fimA*; fimbrial morphology
fimF	528	98	98.08	d	3	Fimbriae; fimbrial morphology
fimG	501	98	98.09	d	3	Fimbriae; fimbrial morphology
fimH	900	98	98.10	d	1	Fimbriae; minor fimbrial subunit, adhesin
firA	1,026	4	4.51	d	1	~ Affects transcription
fis	297	72	73.74	nc	2	Site-specific DNA inversion
flaAI	465	43	43.02	i	1	Flagella; flagellar synthesis and chemotaxis
flaAII	1,005	43	43.03	i	1	Flagella; flagellar synthesis and chemotaxis
flaI	579	42	42.11	d	3	Flagella; flagellar synthesis; regulatory gene
flbB	360	42	42.13	d	3	Flagella; flagellar synthesis; regulatory gene; flagellum-specific sigma factor
fldA	531		15.42	i	2	Flavodoxin
fliD		43	42.70	ns		Flagella; flagellar synthesis, hook-associated protein 2
fliE	315	43	42.72	nc	1	Flagella; basal body structural component
fliF		43	42.71	ns		Flagella; hook-basal body subunit
fnr	753	30	29.99	i	1	Regulatory gene for nitrite and nitrate reductases, hydrogenase, and fumarate reductase
folA	477	1	1.05	d	1	Folate; dihydrofolate reductase (EC 1.5.1.3); trimethoprim resistance
folD	867		11.65	nc	1	Folate; 5,10-methylene-tetrahydrofolate dehydrogenase/5,10-methylyl-tetrahydrofolate cyclohydrolase
fpg	810		82.15	i	1	Fapy-DNA glycosylase
fpp	1,608		80.07	d	2	Dipeptide transport protein
frdA	1,809	94	94.51	i	2	Fumarate reductase (EC 1.3.99.1); flavoprotein subunit
frdB	735	94	94.50	i	2	Fumarate reductase (EC 1.3.99.1); iron-sulfur protein subunit
frdC	360	94	94.48	i	2	Fumarate reductase (EC 1.3.99.1); membrane anchor polypeptide
frdD	396	94	94.49	i	1	Fumarate reductase (EC 1.3.99.1); membrane anchor polypeptide
fruK	939	47	48.11	d	1	Fructose; fructose-1-phosphate kinase (EC 2.7.1.3)
fruR	1,005	2	1.92	d	1	Fructose; regulatory gene; phosphoenolpyruvate:fructose phosphotransferase system repressor
ftsA	1,263	2	2.26	d	1	Cell division; anomalous filamentous growth
ftsE	666	76	77.87	d	1	Cell division; anomalous filamentous growth
ftsQ	831	76	2.24	d	1	Cell division; anomalous filamentous growth
ftsW	1,245		2.14	d	1	Cell division; anomalous filamentous growth
ftsX	1,056	76	77.85	d	1	Cell division; anomalous filamentous growth
ftsY	1,491	76	77.89	d	1	Cell division; anomalous filamentous growth
ftsZ	1,152	2	2.28	d	2	Cell division; anomalous filamentous growth
fucA	645	60	62.45	d	1	Fucose; L-fuculose-1-phosphate aldolase
fucI	1,773	60	62.51	i	1	Fucose; L-fucose isomerase
fucK	1,446	60	62.55	i	1	Fucose; L-fuculose kinase (EC 2.7.1.51)
fucO	1,149	60	62.43	d	1	Fucose; L-1,2-propanediol oxidoreductase
fucP	1,314	60	62.48	i	1	Fucose; L-fucose permease
fucR	729	60	62.59	i	1	Fucose; L-fucose utilization; positive regulatory protein
fucT	189	60	62.61	i	3	Fucose; L-fucose utilization
fucU	399	60	62.58	i	3	Fucose; L-fucose utilization
fumA	1,644	36	36.12	d	1	Fumarate; fumarase
fumB	1,647	93	93.80	i	2	Fumarate; regulatory gene?
fumC	1,401	36	36.08	d	1	Fumarate; fumarase
fur	444	16	15.39	nc	1	Ferric iron uptake; negative regulatory gene
fusA	2,112	73	75.08	d	2	Protein chain elongation factor EF-G
fwd1566	435			nm	1	Biocyclomycin resistance
gabD	1,449	58	59.41	i	1	γ-Aminobutyrate; succinate semialdehyde dehydrogenase (EC 1.2.1.16), NADP-dependent activity
gabP	1,281	58	59.44	i	1	γ-Aminobutyrate; transport of γ-aminobutyrate; GABA transaminase
gabT	1,401	58	59.48	i	1	γ-Aminobutyrate; aminobutyrate aminotransferase (EC 2.6.1.19) activity; GABA permease
galE	1,014	17	17.05	i	1	Galactose; UDP galactose 4-epimerase
galK	1,146	17	17.00	i	1	Galactose; galactokinase (EC 2.7.1.6)
galR	1,032	61	63.36	i	1	Galactose; regulatory gene; repressor of *galETK* operon
galS	1,038		47.67	d	1	Galactose; *mgl* repressor and galactose ultrainduction factor
galT	1,041	17	17.03	i	1	Galactose; galactose-1-phosphate uridyltransferase (EC 2.7.7.12)
gap	993	39	39.95	nc	2	Glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12)
gapB	1,017		65.39	d	1	Glyceraldehyde 3-phosphate dehydrogenase

Continued on following page

TABLE 1—*Continued*

Name	Length (bp)	Map position		*oriC*[c]	Class[d]	Description
		M_G[a]	M_P[b]			
gcvH	390	63	64.85	nc	2	H-protein for glycine cleavage enzyme complex
gdhA	1,344	27	39.14	nc	1	Glutamate dehydrogenase
geneX	939		0.47	d	3	28K protein; function unknown
genF			93.86	ns		Function unknown
genX	978		94.40	i	1	Lysyl-tRNA synthetase homolog
ggt	1,740	76	77.51	d	1	γ-Glutamyltranspeptidase (EC 2.3.2.2)
gidA	1,884	84	84.81	d	1	Glucose-inhibited division; chromosome replication?
gidB	621	84	84.80	d	1	Glucose-inhibited division; chromosome replication?
gldE	2,388		3.10	d	1	Glucose dehydrogenase
glgA	1,434	76	77.12	d	1	Glycogen; glycogen synthase (EC 2.4.1.21)
glgB	2,187	76	77.21	d	1	Glycogen; 1,4-α-glucan branching enzyme (EC 2.4.1.18)
glgC	1,296	76	77.15	d	2	Glycogen; glucose-1-phosphate adenylyltransferase (EC 1.7.7.27)
glgP	2,430		77.06	d	1	Glycogen; α-glucan phosphorylase
glgS	201			nm	3	Glycogen; glycogen synthesis
glgX	1,524	76	77.19	d	1	Glycogen; function unknown
glnA	1,410	88	87.64	i	2	Glutamine; glutamine synthetase (EC 6.3.1.2)
glnB	339		57.03	nc	1	Glutamine; glutamine synthetase; P-II polypeptide
glnG	1,404	88	87.58	i	1	Glutamine; negative regulatory gene for *glnA*
glnH	744	18	18.25	i	2	Glutamine; periplasmic glutamine-binding protein
glnL	1,047	88	87.61	i	1	Glutamine; negative regulatory gene for *glnA*
glnP	657	18	18.23	i	2	Glutamine; glutamine high-affinity transport system; L-glutamine periplasmic binding protein
glnQ	720	18	18.21	i	2	Glutamine; glutamine high-affinity transport system
glnS	1,653	16	15.32	nc	2	Glutamine; glutaminyl-tRNA synthetase (EC 6.1.1.18)
glpA	1,629	49	50.10	i	1	Glycerol phosphate; *sn*-glycerol-3-phosphate dehydrogenase (anaerobic), subunit A
glpB	1,260	49	50.13	i	1	Glycerol phosphate; *sn*-glycerol-3-phosphate dehydrogenase (anaerobic), subunit B
glpC	1,191	49	50.16	i	1	Glycerol phosphate; *sn*-glycerol-3-phosphate dehydrogenase (anaerobic), subunit C
glpD	1,512	75	76.93	d	1	Glycerol phosphate; *sn*-glycerol-3-phosphate dehydrogenase (aerobic)
glpE	393	75	76.97	i	1	Glycerol phosphate; glycogen phosphorylase
glpF	843	89	88.95	i	1	Glycerol phosphate; facilitated diffusion of glycerol
glpG	828	75	76.98	i	1	Glycerol phosphate; glycogen phosphorylase
glpK	1,509	89	88.91	i	2	Glycerol phosphate; glycerol kinase (EC 2.7.1.30)
glpQ	1,074	49	50.10	nc	1	Glycerol phosphate; glycerol-3-phosphate diesterase (EC 3.1.4.2)
glpR	897	75	77.00	i	1	Glycerol phosphate; regulatory gene; glycerol-3-phosphate repressor
glpT	1,356	49	50.07	nc	1	Glycerol phosphate; *sn*-glycerol-3-phosphate permease
gltA		16	16.25	ns		Glutamate; citrate synthase (EC 4.1.3.7)
gltB	4,545	70	72.50	i	1	Glutamate; glutamate synthase, large subunit
gltD	1,416	70	72.60	i	1	Glutamate; glutamate synthase, small subunit
gltI				ns		Glutamate; glutamate decarboxylase
gltP	1,314			nm	1	Glutamate; glutamate and aspartate carrier
gltS	1,206	82	82.65	d	1	Glutamate; glutamate permease
gltX	1,416	52	53.66	d	2	Glutamate; catalytic subunit for glutamyl-tRNA synthetase (EC 6.1.1.17)
gluS	1,827		84.56	d	2	Glucosamine phosphate isomerase
glyA	1,251	55	56.96	d	2	Glycine; serine hydroxymethyl transferase (EC 2.1.2.1)
glySa	912	80	80.47	d	2	Glycine; glycyl-tRNA synthetase, alpha subunit (EC 6.1.1.14)
glySb	2,070	80	80.42	d	2	Glycine; glycyl-tRNA synthetase, beta subunit (EC 6.1.1.14)
gnd	1,407	44	44.90	d	2	Gluconate-6-phosphate dehydrogenase (EC 1.1.1.44), decarboxylating
gor	1,353	77	77.96	nc	1	Glutathionine oxidoreductase (EC 1.6.4.2)
gppA	1,329	85	85.65	i	1	Guanosine pentaphosphatase phosphohydrolase
gpt	459	6	5.68	d	2	Guanine-hypoxanthine phosphoribosyltransferase (EC 2.4.2.8)
greA	475		71.93	d	2	Suppressor gene that restores growth of an RNA polymerase mutant at high temperature
groEL	1,644	94	94.32	d	2	Morphogenesis of phages; head assembly of phages T4 and lambda
groES	291	94	94.31	d	2	Morphogenesis of phages; head assembly of phages T4 and lambda; complementing the *ts*A6 mutation
grpE	591	57	58.86	d	2	Phage lambda replication; host DNA synthesis
grx	258	19	18.83	nc	1	Glutaredoxin
gshI	1,554	58	59.90	d	1	γ-Glutamylcysteine synthetase activity
gshII	948	58	65.81	i	1	Glutathione synthetase (GSH-II) (EC 6.3.2.2)
guaA	1,578	54	55.77	d	2	Guanine; GMP synthetase (EC 6.3.4.1)
guaB	1,467	54	55.80	d	2	Guanine; IMP dehydrogenase (EC 1.2.1.14)
guaC	1,041	3	2.44	nc	1	Guanine; GMP reductase (EC 1.6.6.8)
gutA	1,521	58	60.16	i	1	Sorbitol; D-glucitol-specific enzyme II of phosphotransferase system
gutB	372	58	60.19	i	1	Sorbitol; D-glucitol (sorbitol) specific enzyme III of phosphotransferase system

Continued on following page

TABLE 1—*Continued*

Name	Length (bp)	Map position		*oriC*[c]	Class[d]	Description
		M_G[a]	M_P[b]			
gutD	780	58	60.20	i	1	Sorbitol; glucitol (sorbitol)-6-phosphate dehydrogenase (EC 1.1.1.140)
gutM	360		60.22	i	1	Sorbitol; function unknown
gutQ	669		60.19	nc	1	Sorbitol; function unknown
gutR	774	58	60.23	i	1	Sorbitol; regulatory gene
gyrA	2,634	48	49.78	d	2	Gyrase; DNA gyrase (EC 5.99.1.3), subunit A; resistance or sensitivity to nalidixic acid
gyrB	2,412	83	83.72	d	2	Gyrase; DNA gyrase (EC 5.99.1.3), subunit B; resistance or sensitivity to coumermycin
hag	1,497	43	42.66	d	2	Flagella; flagellar synthesis, filament structural protein; flagellar (H) antigen
helD	2,055		22.02	d	1	DNA; DNA helicase; helicase IV
hemA	1,257	27	27.09	d	1	Hemin; glutamyl-tRNA dehydrogenase
hemB	972	8	8.47	i	1	Hemin; 5-aminolevulinate dehydratase (EC 4.2.1.24) activity
hemC	939	86	86.20	i	1	Hemin; porphobilinogen deaminase (EC 4.3.1.18)
hemD	738	86	86.18	i	1	Hemin; uroporphyrinogen III synthase (EC 4.2.1.75)
hemX	1,179		86.16	i	1	Hemin; urogenIII methylase
heptul	1,041			nm	1	Function unknown
hevA	459		60.68	d	1	Formate hydrogenlyase component
hevB	609		60.67	d	1	Formate hydrogenlyase component
hevC	1,824		60.63	d	1	Formate hydrogenlyase component
hevD	921		60.61	d	1	Formate hydrogenlyase component
hevE	1,707		60.57	d	2	Formate hydrogenlyase component
hevF	540		60.56	d	1	Formate hydrogenlyase component
hevG	765		60.55	d	1	Formate hydrogenlyase component
hevH	408		60.54	d	1	Formate hydrogenlyase component
hfq	309		94.93	d	2	Host factor-I protein for bacteriophage Q beta
himA	300	37	38.39	d	1	Integration host factor, alpha subunit; site-specific recombination
hip	279	20	20.71	d	3	Integration host factor beta subunit; site-specific recombination
hipA	1,323	34	33.32	nc	3	Inhibition of peptidoglycan or DNA synthesis; frequency of persistence following inhibition of murein synthesis
hisA	735	44	44.61	i	1	Histidine; *N*-5-amino-1,4-imidazolecarboxamide isomerase (EC 5.3.1.16)
hisB	1,065	44	44.58	i	1	Histidine; imidazoleglycerol-phosphate dehydratase (EC 4.2.1.19) and histidinol phosphate phosphatase (EC 3.1.3.15)
hisC	1,068	44	44.56	i	1	Histidine; histidinol-phosphate aminotransferase (EC 2.6.1.9)
hisD	1,302	44	44.53	i	1	Histidine; L-histidinol: NAD^+ oxidoreductase (EC 1.1.1.23)
hisF	774	44	44.63	i	1	Histidine; cyclase
hisG	897	44	44.51	i	2	Histidine; ATP phosphoribosyltransferase (EC 2.4.2.17)
hisH	588	44	44.60	i	1	Histidine; amino transferase
hislE	609	44	44.64	i	1	Histidine; phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19); phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.3)
hisL	48		44.50	i	3	Histidine; function unknown
hisM		50	51.62	ns		Histidine; histidine transport
hisP	771	50	51.63	nc	1	Histidine; histidine permease
hisS	1,275	54	55.96	d	1	Histidine; histidyl-tRNA synthetase (EC 6.1.1.21)
hisT	810	50	51.87	d	1	Histidine; pseudouridylate synthetase I
hlyT	489	87	86.94	i	1	Rough; lipopolysaccharide core biosynthesis; positive regulation of production of glucosyltransferase; transcriptional activator of hemolysin synthesis and secretion
hmp	1,191		57.00	nc	1	HMP hemoprotein
hns	414	6	27.70	i	2	DNA-binding protein; histonelike protein HLP-II (HU, BH2, HD, NS)
hsdM	1,587	99	98.76	i	1	Host specificity; host modification; DNA methylase M
hsdR	3,270	99	98.80	i	1	Host specificity; host restriction; endonuclease R
hsdS	1,392	99	98.73	i	3	Host specificity; specificity determinant for *hsdM* and *hsdR*
htpG	1,875	11	10.74	d	2	Heat shock protein C62.5
htpR	855	76	77.83	d	2	RNA polymerase; RNA polymerase (EC 2.7.7.6) σ^{32} subunit; regulatory gene for proteins induced at high temperatures
htpX	882		40.93	d	1	New heat shock protein
htrA	1,473		3.89	d	2	Coding for 51-kDa protein; a σ^{32}-independent mechanism of heat-inducible transcription
htrB	921		23.98	i	1	Function unknown
htrP	759		68.10	nc	2	Function unknown
hupA	270		90.68	nc	2	DNA-binding protein; histonelike protein HU-2
hupB	270	10	10.03	d	2	Histonelike protein HU-1; DNA-binding protein
hyaA	1,119		22.18	d	1	Hydrogen; hydrogenase isoenzyme 1, small subunit
hyaB	1,794		22.21	d	1	Hydrogen; hydrogenase isoenzyme 1, large subunit
hyaC	708		22.25	d	1	Hydrogen; function unknown
hyaD	588		22.26	d	1	Hydrogen; function unknown
hyaE	399		22.27	d	1	Hydrogen; function unknown

Continued on following page

TABLE 1—*Continued*

Name	Length (bp)	Map position		*oriC*[c]	Class[d]	Description
		M.G[a]	M.P[b]			
hyaF	858		22.28	d	1	Hydrogen; function unknown
hydG	1,314		90.74	nc	1	Hydrogenase activity
hydH			90.73	ns		Hydrogenase activity
hypA	351		60.68	i	3	Hydrogen; hydrogenase isoenzyme
hypB	873		60.69	i	1	Hydrogen; hydrogenase isoenzyme
hypC	273		60.71	i	1	Hydrogen; hydrogenase isoenzyme
hypD	1,122		60.71	i	1	Hydrogen; hydrogenase isoenzyme
hypE	969		60.74	i	1	Hydrogen; hydrogenase isoenzyme
iap	1,038	59	61.27	i	3	Altered isozyme pattern of alkaline phosphatase
icd	1,251	26	25.60	d	2	Isocitrate dehydrogenase, $NADP^+$ specific (EC 1.1.1.42)
iciA	894		64.91	nc	1	Chromosome initiation inhibitor
iclR	825	91	91.17	i	1	Acetate; regulatory gene for *aceBA* operon; repressor protein
ileR	303	100	99.85	nc	1	Isoleucine; regulatory gene; negative regulatory of *thr* and ilv operons
ileS		0	0.56	ns		Isoleucine; isoleucyl-tRNA synthetase (EC 6.1.1.5)
ilvA	1,542	85	85.47	d	1	Isoleucine-valine; threonine deaminase (EC 4.2.1.16)
ilvB	1,686	83	83.15	d	1	Isoleucine-valine; acetohydroxy acid synthase I (EC 4.1.3.18), valine-sensitive, large subunit
ilvC	1,473	85	85.52	d	2	Isoleucine-valine; ketol-acid reductoisomerase (EC 1.1.1.86)
ilvD	1,707	85	85.43	d	2	Isoleucine-valine; dihydroxyacid dehydratase (EC 4.2.1.9)
ilvE	927	85	85.41	d	2	Isoleucine valine; branched-chain amino acid aminotransferase (EC 2.6.1.42)
ilvG	1,644	85	85.36	d	1	Isoleucine-valine; acetohydroxy acid synthase II (EC 4.1.3.18) valine insensitive, large subunit
ilvH	492	2	1.90	d	1	Isoleucine-valine; acetohydroxy acid synthase III (EC 4.1.3.18) valine sensitive, small subunit
ilvI	1,701	2	1.87	d	1	Isoleucine-valine; acetohydroxy acid synthase II (EC 4.1.3.18) valine insensitive, large subunit
ilvL	96		85.36	d	3	Isoleucine-valine; function unknown
ilvM	258	85	85.40	d	3	Isoleucine-valine; acetohydroxy acid synthase II (EC 4.1.3.18), valine insensitive, small subunit
ilvN	288	83	83.14	d	1	Isoleucine-valine; acetohydroxy acid synthase II (EC 4.1.3.18) valine sensitive, small subunit
ilvY	894	85	85.50	i	1	Isoleucine-valine; positive regulatory gene for *ilvC*
infA	183	20	19.92	i	3	Protein chain initiation factor 1 (IF1)
infB	2,673	69	71.61	d	2	Protein chain initiation factor 2
infC	543	38	38.50	d	1	Protein chain initiation factor 3
IS*3*	1,164		12.24	i	3	Insertion sequence; integrase
IS*30*	1,149			nm	3	Insertion sequence; transposase
ispA	900			nm	1	Farnesyl diphosphate synthase (EC 2.5.1.1)
katE	2,262	38	38.80	i	1	Catalase; biosynthesis of catalase hydroperoxidase HPII
katF	1,086	59	61.05	nc	1	Catalase; biosynthesis of catalase hydroperoxidase HPII (III) and exonuclease III; regulatory gene
katG	2,181	89	89.28	d	2	Catalase; catalase-peroxidase hydroperoxidase HPI (I), structural gene
kbl	1,194	81	81.87	d	1	2-Amino-3-ketobutyrate-CoA ligase (EC 2.3.1.29) (glycine acetyltransferase)
kdpA	1,674	16	15.71	i	1	Potassium dependence; high-affinity potassium transport system; probably K^+-stimulated ATPase
kdpB	2,049	16	15.67	i	1	Potassium dependence; high-affinity potassium transport system
kdpC	573	16	15.66	i	1	Potassium dependence; high-affinity potassium transport system
kdpD	2,685	16	15.60	i	1	Potassium dependence; high-affinity potassium transport system; regulatory gene
kdpE	678		15.59	i	1	Potassium dependence; high-affinity potassium transport system; cytoplasmic protein
kdsA	855	27	27.18	d	2	3-Deoxy-D-manno-octulosonic acid 8-phosphate synthase
kdsB	747	85	85.21	nc	1	CTP:CMP-3-deoxy-D-manno-octulosonate cytidylyltransferase
kdtA	1,278		82.18	d	1	KDO transferase
kefC	1,863	1	1.01	nc	1	K^+ efflux; NEM-activable K^+/H^+ antiporter; glutathione regulated potassium efflux system
ksgA	819	1	1.10	i	1	Kasugamycin; *S*-adenosylmethionine-6-*N'*,*N'*-adenosyl dimethyltransferase (16S rRNA)
lacA	612	8	7.89	i	3	Lactose; galactoside acetyltransferase (EC 2.3.1.18)
lacI	1,083	8	8.00	i	1	Lactose; regulatory gene; repressor protein of *lac* operon
lacY	1,254	8	7.90	i	1	Lactose; galactoside permease (M protein)
lacZ	3,075	8	7.93	i	1	Lactose; β-D-galactosidase (EC 3.2.1.23)
lamB	1,338	92	91.69	d	2	Lambda; phage lambda receptor protein; maltose high-affinity uptake system
lepA	1,797	55	57.39	d	2	GTP-binding membrane protein; function unknown
lepB	972	55	57.37	d	1	Leader peptidase (signal peptidase I)
leuA		2	1.82	ns		Leucine; α-isopropylmalate synthase (EC 4.1.3.12)
leuO	1,050		1.84	d	3	Leucine; function unknown
leuS	2,580	15	14.60	i	2	Leucine; leucyl-tRNA synthetase (EC 6.1.1.14)

Continued on following page

TABLE 1—*Continued*

Name	Length (bp)	Map position		*oriC*[c]	Class[d]	Description
		M.G[a]	M.P[b]			
lexA	606	92	91.77	nc	1	Resistance or sensitivity to X rays and UV; regulatory gene of SOS operon
lig	2,016	52	53.86	d	1	DNA; DNA ligase
lip	846	15	15.09	nc	2	Lipoate; synthesis of α-(+)-lipoic acid; lipoic acid synthetase
lit	894	25	25.70	nc	3	Phage T4 late gene expression; locus of element e14
livF	714	76	77.68	d	1	Leucine, isoleucine, and valine; high-affinity branched-chain amino acid transport system
livG	768	76	77.69	d	1	Leucine, isoleucine, and valine; high-affinity branched-chain amino acid transport system; membrane component
livH	927	76	77.73	d	1	Leucine, isoleucine, and valine; high-affinity branched-chain amino acid transport system; membrane component
livJ	1,104	76	77.80	d	2	Leucine, isoleucine, and valine; periplasmic binding protein; high-affinity branched-chain amino acid transport system
livK	1,110	76	77.76	d	1	Leucine, isoleucine, and valine; leucine-specific periplasmic binding protein; high-affinity branched-chain amino acid transport system
livM	1,275	76	77.71	d	1	Leucine, isoleucine, and valine; high-affinity branched-chain amino acid transport system
livR	501	20	20.13	nc	1	Leucine, isoleucine, and valine; high-affinity branched-chain amino acid transport system; regulatory gene; repressor protein
lpd	1,425	3	2.76	d	2	Lipoamide dehydrogenase (NADH) (EC 1.6.4.3)
lppX	234	36	37.82	nc	2	Murein lipoprotein structural gene
lpxA	789	4	4.55	d	1	UDP-*N*-acetylglucosamine acetyltransferase; lipid A biosynthesis protein
lpxB	1,149	4	4.56	d	1	Lipid A disaccharide synthase
lrp	495		20.08	nc	3	Leucine-responsive regulatory protein
lsp	495	1	0.57	d	1	Prolipoprotein signal peptidase (SPaseII)
luxH	654		68.63	d	1	Function unknown
lysA	1,263	61	63.38	d	1	Lysine; diaminopimelate decarboxylase (EC 4.1.1.20)
lysC	1,350	91	91.34	i	1	Lysine; aspartokinase III
lysP	1,470		47.82	d	2	Lysine; lysine specific permease
lysR	936	61	63.41	i	1	Lysine; regulatory gene; activator of *lysA*
lysS	1,518	62	64.59	d	2	Lysine; lysyl-tRNA synthetase constitutive; suppressor of ColE1 mutation in primer RNA
lysU	1,506	94	93.93	i	1	Lysine; lysyl-tRNA synthetase, inductible
malE	1,188	92	91.63	i	2	Maltose; periplamic maltose-binding protein; subtrate recognition for transport and chemotaxis
malF	1,542	92	91.60	i	1	Maltose; maltose transport; cytoplasmic membrane protein
malG	885	92	91.58	i	2	Maltose; active transport of maltose and maltodextrins
malI	975	36	36.34	nc	1	Maltose; production of oligosaccharide, probably glucose polymer
malJ	978		36.37	d	1	Maltose; function unknown
malK	1,179	92	91.66	d	1	Maltose; maltose permeation
malP	2,391	75	76.72	d	1	Maltose; maltodextrin phosphorylase (EC 2.4.1.1)
malQ	2,085	75	76.67	d	1	Maltose; amylomaltase (EC 2.4.1.25)
malS	2,031	80	80.73	i	1	Maltose; α-amylase precursor (EC 3.2.1.1)
malT	2,706	75	76.78	i	1	Maltose; positive regulatory gene for *mal* regulon
malX	1,593		36.39	i	1	Maltose; function unknown
malY	1,173		36.43	i	1	Maltose; function unknown
malZ	1,815		9.20	d	1	Maltose; α-1,4-D-glucosidase (EC 3.2.1.20)
manA	1,176	36	36.14	i	1	Mannose; mannose-6-phosphate isomerase (EC 5.3.1.8)
map	795		4.20	nc	1	Methionine amino peptidase
mcrA	834	25	25.94	nc	3	Restriction of DNA at 5-methyl cytosine residues; locus of e14
mcrB	1,395	98	98.68	i	3	Restriction of DNA at 5-methylcytosine residues
mcrC	1,074	98	98.66	i	3	Cytosine-specific endonuclease
mcrD	948		98.63	nc	1	Restriction of DNA; function unknown
mdh	936	70	73.11	d	2	Malate dehydrogenase (EC 1.1.1.37)
melA	1,353	93	93.71	d	1	Melibiose; α-galactosidase (EC 3.2.1.22)
melB	1,410	93	93.74	d	1	Melibiose; thiomethyl galactoside permease II
menB	858	49	50.57	d	2	Menaquinone; 1,4-dihydroxy-2-naphthoate (DHNA) synthase
menD	1,386	49	50.64	d	1	Menaquinone; menaquinone biosynthesis; 2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase
mepA	822	50	52.11	d	1	Penicillin-insensitive murein DD-endopeptidase
metA	927	91	90.98	nc	1	Methionine; homoserine transsuccinylase (EC 2.3.1.46)
metB	1,161	89	89.18	d	1	Methionine; cystathionine γ-synthase (EC 4.2.99.9)
metC	1,188	65	67.94	i	1	Methionine; cystathionine γ-lyase (EC 4.4.1.1)
metE		86	86.69	ns		Methionine; tetrahydropteroyltriglutamate methyltransferase (EC 2.1.1.14)
metF	888	89	89.25	d	1	Methionine; 5,10-methylenetetrahydrofolate reductase (EC 1.1.1.68)
metG	2,034	46	46.67	i	2	Methionine; methionyl-tRNA synthetase
metH	3,600	91	91.19	d	1	Methionine; vitamin B_{12}-dependent homocysteine-N^5-methyl tetrahydrofolate transmethylase

Continued on following page

TABLE 1—*Continued*

Name	Length (bp)	Map position		*oriC*[c]	Class[d]	Description
		M.G[a]	M.P[b]			
metJ	456	89	89.17	i	1	Methionine; regulatory gene : repressor of *metF*
metK	1,155	64	65.71	i	2	Methionine; methionine adenosyltransferase (EC 2.5.1.6)
metL	2,433	89	89.21	d	1	Methionine; aspartokinase II (EC 2.7.2.4), homoserine dehydrogenase II (EC 1.1.1.3)
metR	954	86	86.67	nc	1	Methionine; regulatory gene of *metE* and *metH*
mglA		46	47.63	ns		Methylgalactoside; methylgalactoside transport and galactose taxis; cytoplasmic membrane protein
mglB	999	46	47.61	i	2	Methylgalactoside; galactose-binding protein; receptor for galactose taxis
mglC	1,011	46	47.66	i	1	Methylgalactoside; methyl-galactoside transport and galactose taxis
miaA	951	95	94.91	d	1	Δ^2-isopentenyl PP_i transferase; 2-methylthio-N^6-isopentyladenosine hypermodification
minC	696	26	26.26	i	1	Formation of minute cells containing no DNA; proper placement of the division septum
minD	813	26	26.25	i	1	Formation of minute cells containing no DNA; proper placement of the division septum
minE	267	26	26.24	i	1	Formation of minute cells containing no DNA; proper placement of the division septum
molA	918	92	91.73	d	1	Maltose; probably for an exported protein; periplasmic protein
motA	888	42	42.12	d	1	Motility; chemotactic response; flagellar paralysis
motB	927	42	42.10	d	1	Motility; chemotactic response; flagellar paralysis
mprA	531		59.91	nc	1	This gene seems to be involved in the control of microcin B17 synthesis
mreB		71	73.48	ns		Cell shape; sensitivity to antibiotics
mreC	1,104	71	73.47	d	1	Cell shape; sensitivity to antibiotics
mreD	489	71	73.46	d	1	Cell shape; sensitivity to antibiotics
mrp	1,110		46.65	d	1	putative ATPase
mrr	912	99	98.87	d	1	Restriction of methylated adenine
msbB	972		41.41	nc	1	Multicopy suppressor of null mutations in the high-temperature requirement gene *htrB*; membrane-bound lytic transglycosylase
msyB	372		23.94	i	3	Multicopy suppressor of *secY24* mutation
mtlA	1,911	81	81.46	i	2	Mannitol; mannitol-specific enzyme II of phosphotransferase system
mtlD	1,059	81	81.51	i	3	Mannitol; mannitol-1-phosphate dehydrogenase (EC 1.1.1.17)
mtr	1,245	69	71.43	d	1	Methyltryptophan; resistance to 5-methyltryptophan; tryptophan-specific permease
mukB	4,605		20.95	d	1	Protein involved in chromosome partitioning
murC	1,476	2	2.19	d	1	Murein; UDP-*N*-acetylmuramate : L-alanine ligase
murD	1,317	2	2.11	d	1	Murein; UDP-*N*-acetylmuramoyl-L-alanine; D-glutamate ligase
murE	1,488	2	2.03	d	1	Murein; *meso*-diaminopimelate-adding enzyme; UDP-MurNac-tripeptide synthetase
murF	1,359	2	2.06	d	1	Murein; D-alanyl:D-alanine adding enzyme
murG	1,044	2	2.17	d	1	Murein; murein or envelope biosynthesis
murX	1,083		2.09	d	1	Murein; function unknown
mutD	729	5	5.26	d	1	Mutator and DNA; DNA polymerase III holoenzyme, epsilon subunit
mutH	687	61	63.22	nc	3	Mutator; increased rates of frameshift and base substitution mutations; methyl-directed mismatch repair
mutL	1,848	95	94.87	d	1	Mutator; methyl-directed mismatch repair
mutS	2,562	59	60.83	i	1	Mutator; methyl-directed mismatch repair
mutT	390	2	2.41	d	1	Mutator; high rate of AT-GC transversions
mutY	1,053	64	66.05	i	1	Mutator; adenine glycosylase; GC→TA transversions
mvrA	807	7	6.91	nc	1	Methyl viologen resistance
mvrC	333		12.29	nc	3	Membrane protein (predicted by hydropathy); ethidium bromide efflux; ethidium bromide resistance
nadA	840	17	16.85	nc	3	NAD; quinolinate synthetase, A protein
nadB	1,620	56	57.57	i	1	NAD; quinolinate synthetase, B protein
nagA	1,149	16	15.25	i	1	*N*-Acetylglucosamine; *N*-acetylglucosamine-6-phosphate deacetylase (EC 3.5.1.25)
nagB	801	16	15.28	i	2	*N*-Acetylglucosamine; glucosamine-6-phosphate deaminase
nagC	1,221	16	15.22	i	1	*N*-Acetylglucosamine; function unknown
nagD	753	16	15.21	i	1	*N*-Acetylglucosamine; function unknown
nagE	1,947	16	15.30	d	2	*N*-Acetylglucosamine; *N*-acetylglucosamine-specific enzyme II of phosphotransferase system
narG	3,717	27	27.43	d	2	Nitrate reductase; nitrate reductase (EC 1.7.99.4), alpha subunit
narH	1,536	27	27.51	d	2	Nitrate reductase; nitrate reductase (EC 1.7.99.4), beta subunit
narI	675	27	27.56	d	2	Nitrate reductase; nitrate reductase (EC 1.7.99.4), gamma subunit; cytochrome *b*NR, structural gene
narJ	708	27	27.55	d	1	Nitrate reductase; nitrate reductase (EC 1.7.99.4), delta subunit
narK	1,386	27	27.39	d	1	Nitrate reductase; regulatory gene
narL	648	27	27.34	i	1	Nitrate reductase; regulatory gene

Continued on following page

TABLE 1—*Continued*

Name	Length (bp)	Map position		*oriC*[c]	Class[d]	Description
		M.G[a]	M.P[b]			
narU	1,386		33.05	d	1	Nitrate reductase; function unknown
narV	678		32.91	i	1	Nitrate reductase; function unknown
narW	693		32.92	i	1	Nitrate reductase; function unknown
narX	1,794	27	27.35	i	1	Nitrate reductase; regulatory gene
narY	1,542		32.94	i	1	Nitrate reductase; function unknown
narZ	3,774	33	32.97	i	1	Nitrate reductase; cryptic gene(s) encoding a second nitrate reductase
ndh	1,302	22	25.10	d	1	Respiratory NADH dehydrogenase
nfo	858	47	47.89	i	1	Endonuclease IV
nirB	2,517	74	75.78	nc	2	Nitrate reductase; NADH-nitrate oxidoreductase (EC 1.6.6.4) apoprotein, structural gene
nirC	807	26	75.84	nc	1	Nitrate reductase; NADH-nitrite reductase (EC 1.6.6.4) activity
nirD	324		75.83	nc	1	Nitrate reductase; NADH-nitrite reductase (EC 1.6.6.4) activity
nlpA	819			nm	1	Lipoprotein-28
nlpB	429		55.05	d	2	Lipoprotein-34
nmpC	1,125	13	12.46	d	2	New membrane protein; production of an outer membrane porin protein
npl	891		72.88	nc	1	N-acetylneuraminate lyase
nrdA	2,331	49	49.92	i	1	Ribonucleoside diphosphate reductase (EC 1.17.4.1); subunit B1
nrdB	1,131	49	49.97	i	2	Ribonucleoside diphosphate reductase (EC 1.17.4.1); subunit B2
nth	636	36	36.81	nc	1	"Endonuclease III"; DNA glycosylase and phosphoric monoester lyase
ntr	825		39.01	nc	1	Nitrogen fixation
nupG	1,254	64	66.11	i	2	Transport of nucleosides
nusA	1,485	69	71.67	d	2	Transcription termination; L factor
nusB	420	10	10.40	i	2	Transcription termination; L factor
nusG	546		90.21	d	2	Transcription termination; L factor
ogr	216			nm	3	Positive regulator of phage P2 late gene transcription
ogt	513		30.01	i	1	DNA repair; O^6-alkylguanine-DNA-alkyltransferase
ompA	1,038	22	21.90	i	2	Outer membrane protein; outer membrane protein 3a, structural gene
ompC	1,104	48	49.28	d	2	Outer membrane protein; outer membrane protein 1b, structural gene
ompF	1,089	21	21.18	i	2	Outer membrane protein; outer membrane protein 1a, structural gene
ompR	855	75	76.45	d	1	Outer membrane protein; positive regulatory gene for OmpC and OmpF
ompT	951	13	12.66	i	3	Outer membrane protein; outer membrane protein 3b, a protease
oppA	1,632	28	27.88	nc	1	Oligopeptide transport; periplasmic binding protein
osmB	219		28.83	nc	1	Lipoprotein; osmotically inducible protein
osmC	417		32.77	nc	1	Osmotically inducible protein
oxyR	918		89.79	d	1	Morphology and autoaggregation control protein
P-14	387		99.18	i	3	Function unknown
P-18	498		99.14	i	1	Function unknown
pabA	564	74	75.46	d	1	*p*-Aminobenzoate; *p*-aminobenzoate synthetase, CoI
pabB	1,362	40	40.58	nc	1	*p*-Aminobenzoate; *p*-aminobenzoate synthetase, CoII
pabC	810			nm	3	*p*-Aminobenzoate; 4-amino-4-deoxychorismate lyase
pal	519		16.80	nc	2	Peptidoglycan-associated lipoprotein
panF	1,449	71	73.64	i	1	Pantothenate; pantothenate permease
parC	2,193		68.19	d	1	DNA; topoisomerase IV subunit
parE	1,806		68.41	d	1	DNA; topoisomerase IV subunit
patA	1,137	89	15.96	nc	1	Putrescine aminotransferase activity; transport protein
patB	828	89	15.98	nc	1	Putrescine aminotransferase activity; transport protein
patC	795	89	16.00	nc	1	Putrescine aminotransferase activity; transport protein
patD	1,047	89	16.02	nc	2	Putrescine aminotransferase activity; transport protein
patE	1,320		15.50	i	1	Putrescine transport protein
pbpA	1,899	15	14.46	i	1	Cell shape; penicillin-binding protein 2
pbpB	1,767	2	1.99	d	1	Peptidoglycan synthetase; septum formation; penicillin-binding protein 3
pckA	1,395	75	76.38	i	2	Phosphoenolpyruvate carboxykinase (EC 4.1.1.49)
pcm	627		61.10	nc	1	L-Isoaspartyl protein carboxyl methyltransferase type II
pdxA	987	1	1.12	i	1	Pyridoxine; requirement
pdxB	1,137	50	51.91	d	1	Pyridoxine; placement of 5,5′ and 6′ carbons into pyridine ring of pyridoxine
pdxJ	732	56	57.30	d	1	Pyridoxine; requirement; pyridoxal phosphate biosynthesis
pepD	1,458	6	5.66	nc	1	Peptides; peptidase D, a dipeptidase
pepN	2,613	21	21.29	d	1	Peptides; aminopeptidase N
pepP	1,326		64.31	nc	1	Peptides; proline amino peptidase II
pepQ	1,938		87.09	d	1	Peptides; proline dipeptidase
pfkA	960	89	88.72	d	2	6-phosphofructokinase I (EC 2.7.1.11)
pfkB	927	38	38.63	i	1	Level of 6-phosphofructokinase II; suppressor of *pfkA*
pfl	2,280	20	20.44	i	2	Pyruvate formate lyase
pflCo	738	20	20.42	i	1	Pyruvate formate-lyase-activating enzyme
pfs	660		3.84	i	2	Function unknown
pga	2,541	31	31.68	i	3	Penicillin G acylase; phenylacetate degradation
pgi	1,647	91	91.38	d	2	Glucose phosphate isomerase (EC 5.3.1.9)

Continued on following page

TABLE 1—*Continued*

Name	Length (bp)	Map position		*oriC*[c]	Class[d]	Description
		M.G[a]	M.P[b]			
pgk	1,161	63	65.37	d	2	Phosphoglycerate kinase (EC 2.7.2.3)
pgm		15	15.05	ns		Phosphoglucomutase (EC 2.7.5.1)
pgpA	504	10	9.69	nc	3	Phosphatidylglycerophosphate phosphatase, membrane bound
pgsA	651	42	42.45	d	1	Phosphatidylglycerophosphate synthetase (EC 2.7.8.5)
pheA	1,158	57	58.17	nc	1	Phenylalanine; chorismate mutase-P-prephenate dehydrogenase
pheP	1,377	13	12.81	nc	1	Phenylalanine; associated with the phenylalanine-specific transport system; permease
pheS	996	37	38.45	d	2	Phenylalanine; phenylalanyl-tRNA synthetase, alpha subunit (EC 6.1.1.20)
pheT	2,388	37	38.40	d	2	Phenylalanine; phenylalanyl-tRNA synthetase, beta subunit (EC 6.1.1.20)
phnA	336	92	93.37	i	2	Alkylphosphonate uptake; *psiD* locus
phnB	444	92	93.35	i	1	Alkylphosphonate uptake; *psiD* locus
phnC	789	92	93.33	i	1	Alkylphosphonate uptake; *psiD* locus
phnD	1,017	92	93.31	i	1	Alkylphosphonate uptake; *psiD* locus; carbon-phosphorus lyase
phnE	831	92	93.29	i	1	Alkylphosphonate uptake; *psiD* locus
phnF	726	92	93.27	i	1	Alkylphosphonate uptake; *psiD* locus
phnG	453	92	93.26	i	1	Alkylphosphonate uptake; *psiD* locus
phnH	585	92	93.25	i	1	Alkylphosphonate uptake; *psiD* locus
phnI	1,065	92	93.23	i	1	Alkylphosphonate uptake; *psiD* locus
phnJ	846	92	93.21	i	1	Alkylphosphonate uptake; *psiD* locus
phnK	759	92	93.19	i	1	Alkylphosphonate uptake; *psiD* locus
phnL	681	92	93.18	i	1	Alkylphosphonate uptake; *psiD* locus
phnM	1,137	92	93.15	i	1	Alkylphosphonate uptake; *psiD* locus
phnN	558	92	93.14	i	1	Alkylphosphonate uptake; *psiD* locus
phnO	435	92	93.13	i	1	Alkylphosphonate uptake; *psiD* locus
phnP	759	92	93.12	i	1	Alkylphosphonate uptake; *psiD* locus
phnQ	366	92	93.11	i	1	Alkylphosphonate uptake; *psiD* locus
phoA	1,416	9	8.74	d	1	Phosphate; alkaline phosphatase (EC 3.1.3.1)
phoB	687	9	9.10	d	1	Phosphate; positive regulatory gene for *pho* regulon
phoE	1,053	6	5.73	i	1	Phosphate; outer membrane protein e, structural gene
phoM	1,425	100	99.89	d	1	Phosphate; positive regulatory gene for *pho* regulon
phoP	672		25.50	i	1	Phosphate; regulator protein
phoQ	1,461		25.46	i	1	Phosphate; sensor protein
phoR	1,293	9	9.10	d	1	Phosphate; positive and negative regulatory gene for *pho* regulon
phoS	1,041	84	84.53	d	2	Phosphate; periplasmic phosphate-binding protein
phoU	726	84	84.45	d	2	Phosphate; high-affinity phosphate-specific transport system; regulatory gene
phoW	960	84	84.51	d	2	Phosphate; periplasmic phosphate-binding protein; high-affinity phosphate-specific transport system
phr	1,419	16	16.07	nc	1	Photoreactivation; deoxyribodipyrimidine photolyase (EC 4.1.99.3)
pilin	852			nm	3	Pili?
pin	552	26	25.93	d	3	Inversion of adjacent DNA; invertible-P region of the excisable element e14
pk-1	1,389			nm	2	Pyruvate kinase I
pldA	870	86	86.52	d	1	Detergent-resistant phospholipase A activity
pldB	1,020	86	86.70	nc	1	Lysophospholipase L2
plsB	2,424	92	91.81	i	1	Phospholipid synthesis; glycerolphosphate acyltransferase activity
plsC	738		68.17	d	1	Phospholipid synthesis; 1-acylglycerol-3-phosphate acyltransferase
pmbA	1,353		96.16	nc	1	Involved in the production of antibiotic MccB17
pmi	1,371		15.01	i	1	Mannose; phosphomanose isomerase
pncB	1,203	21	21.25	i	1	Pyridine nucleotide cycle; nicotinate phosphoribosyltransferase (EC 2.4.2.11)
pnp	2,136	69	71.52	d	2	Polynucleotide phosphorylase (EC 2.7.7.8)
pntA	1,506	35	35.91	d	1	Pyridine nucleotide transhydrogenase (EC 1.6.1.1), alpha subunit
pntB	1,386	35	35.88	d	1	Pyridine nucleotide transhydrogenase (EC 1.6.1.1), beta subunit
polA	2,787	87	87.44	d	1	Polymerase; DNA polymerase I (EC 2.7.7.7)
polB	2,307	2	1.35	i	1	Polymerase; DNA polymerase II (EC 2.7.7.7)
ponA	2,550	75	76.17	i	1	Murein; peptidoglycan synthetase; cell wall synthesis; penicillin-binding protein 1A
ponB	2,532	75	3.54	d	1	Murein; peptidoglycan synthetase; cell wall synthesis; penicillin-binding protein 1Bs
popC	1,281	4	3.72	i	1	Porphyrin; synthesis of δ-aminolevulinate; glutamate-1-semialdehyde aminotransferase
poxB	1,716	19	19.58	nc	1	Pyruvate oxidase (EC 1.2.2.2), structural gene; cytochrome *b* oxidoreductase
ppc	2,649	89	89.62	i	1	Phosphoenolpyruvate; phosphoenolpyruvate carboxylase (EC 4.1.1.31)
ppfA	627			nm	2	Required for the formation of correctly folded alcaline phosphatase; formation of disulfide bridges
ppiA	573		75.49	d	2	Peptidyl-prolyl *cis-trans* isomerase a
ppiB	495		12.01	nc	2	Peptidyl-prolyl *cis-trans* isomerase
pps	2,382	37	38.18	d	2	Phosphoenolpyruvate; phosphoenolpyruvate synthase (EC 2.7.9.2)

Continued on following page

TABLE 1—*Continued*

Name	Length (bp)	Map position		*oriC*[c]	Class[d]	Description
		M.G[a]	M.P[b]			
prc	2,049		40.96	d	1	Involved in the C-terminal processing of penicillin-binding protein 3; tail-specific protease
prfA	972	27	27.12	d	1	Protein release factor 1
prfB	804	62	64.62	d	2	Protein release factor 2
priA	2,199		89.08	i	1	Prisomal protein n′
priC	528		10.62	nc	1	Prisomal replication protein n′
prlF	336		70.98	nc	1	Suppressor of the *htrA* null phenotype; protein export?
proA	1,221	6	5.78	d	1	Proline; γ-glutamyl phosphate reductase (EC 1.2.1.41)
proB	1,101	6	5.76	d	1	Proline; γ-glutamyl kinase (EC 2.7.2.11)
proC	810	9	8.82	d	1	Proline; pyrroline-5-carboxylate reductase (EC 1.5.1.2)
proS	1,551	5	4.85	i	2	Proline; prolyl-tRNA synthetase (EC 1.1.1.15); DNA and RNA biosynthesis factor
proV	1,203	57	59.68	i	1	Proline; High-affinity transport system for glycine betaine and proline
proW	1,065	57	59.70	i	1	Proline; High-affinity transport system for glycine betaine and proline
proX	993		59.72	i	1	Proline; function unknown
prrB	1,203	31	30.82	i	3	γ-Aminobutyraldehyde (pyrroline) dehydrogenase activity
prrC	1,188	31	30.79	i	3	γ-Aminobutyraldehyde (pyrroline) dehydrogenase activity
prrD	939	31	30.77	i	1	γ-Aminobutyraldehyde (pyrroline) dehydrogenase activity
prs	945	26	27.03	i	2	Phosphoribosyl pyrophosphate synthetase (EC 2.7.6.1)
psd	969	95	94.70	i	2	Phosphatidylserine decarboxylase
pspA	669		31.04	nc	1	Stress-induced *psp* operon
pspB	225		31.06	nc	1	Stress-induced *psp* operon
pspC	360		31.06	nc	1	Stress-induced *psp* operon
pspD	222		31.07	nc	3	Stress-induced *psp* operon
pspE	315		31.08	nc	1	Stress-induced *psp* operon
pss	1,359	56	57.78	nc	1	Phosphatidylserine synthetase (EC 2.7.8.8)
pstA	891	84	84.49	d	2	High-affinity phosphate-specific transport system
pstB	774	84	84.47	d	2	High-affinity phosphate-specific transport system; cytoplasmic membrane protein?
pth	585	26	26.79	nc	1	Peptidyl-tRNA hydrolase
ptr	2,886	61	62.93	d	1	Protease III
ptsG	1,434	25	24.92	d	2	Phosphotransferase system; glucose phosphotransferase enzyme II
ptsH	258	52	53.99	d	2	Phosphotransferase system; phosphohistidinoprotein-hexose phosphotransferase (EC 2.7.1.69)
ptsI	1,728	52	53.95	d	2	Phosphotransferase system; phosphotransferase system enzyme I
ptsL	972	40	40.72	i	1	Mannose; mannose phosphotransferase enzyme III-Man; permease
ptsM	861	40	40.76	i	2	Mannose; mannose phosphotransferase enzyme II-M-Man; permease
ptsP	801	40	40.74	i	2	Mannose; mannose phosphotransferase enzyme II-P-Man; permease; penetration of phage lambda
purA	1,299	95	95.03	nc	2	Purine; adenylosuccinate synthetase (EC 6.3.4.4)
purB	1,371	25	25.51	i	2	Purine; adenylosuccinate lyase (EC 4.3.2.2)
purC	711	53	55.03	d	2	Purine; 5′-phosphoribosyl-5-aminoimidazole-4-*N*-succinocarboxamide synthetase (EC 6.3.2.6)
purD	1,287	90	90.78	i	1	Purine; phosphoribosyl glycineamide synthetase (EC 6.3.4.13)
purE	510	12	11.96	d	2	Purine; 5′-phosphoribosyl-5-amino-4-imidazole carboxylase I (EC 4.1.1.21), catalytic subunit
purF	1,515	50	51.74	d	1	Purine; amidophosphoribosyl transferase (EC 2.4.2.14)
purH	1,587	90	90.81	i	2	Purine; phosphoribosyl aminoimidazole carboxamide formyltransferase (EC 2.1.2.3)
purK	1,068	12	11.97	d	1	Purine; 5′-phosphoribosyl-5-amino-4-imidazole carboxylase II (EC 4.1.1.21), CO2-fixing subunit
purL	3,888	55	57.11	d	2	Purine; phosphoribosyl formylglycine amide synthetase (EC 6.3.5.3); homologous to *purG* of *S. typhimurium*
purM	1,038	54	55.56	i	1	Purine; 5′-phosphoribosyl-5-aminoimidazole synthetase (EC 6.3.3.1); homologous to *purI* of *S. typhimurium*
purN	639	54	55.58	i	1	Purine; 5′-phosphoribosyl glycinamide transformylase (EC 2.1.2.2)
purR	1,023	36	37.33	nc	1	Purine; purine nucleotide synthesis repressor protein; regulatory gene of *pur* regulon
putA		23	23.18	ns		Proline utilization; proline dehydrogenase (EC 1.5.99.8)
putP	1,506	23	23.19	d	1	Proline utilization; major proline permease
pykA	1,443		41.19	nc	2	Pyruvate kinase type II
pyrB	936	97	96.34	i	1	Pyrimidine; aspartate carbamoytransferase (EC 2.1.3.2) catalytic subunit
pyrC	1,044	24	24.10	i	1	Pyrimidine; dihydroorotase (EC 3.5.2.3)
pyrD	1,008	21	21.58	d	1	Pyrimidine; dihydroorotate oxidase (EC 1.3.3.1)
pyrE	636	82	82.36	d	1	Pyrimidine; orotate phosphoribosyl transferase (EC 2.4.2.10)
pyrF	738	28	28.79	d	1	Pyrimidine; orotidine-5′-phosphate decarboxylase (EC 4.1.1.23)
pyrG	1,638	60	61.93	d	2	Pyrimidine; CTP synthetase (EC 6.3.4.2)
pyrI	462	97	96.33	i	1	Pyrimidine; aspartate carbamoyltransferase (EC 2.1.3.1.2), regulatory subunit

Continued on following page

TABLE 1—*Continued*

Name	Length (bp)	Map position		*oriC*[c]	Class[d]	Description
		M.G[a]	M.P[b]			
que-1	345		9.23	i	1	Queuosin biosynthesis
que-2	1,071		9.25	d	1	Queuosin biosynthesis
racC	276	30	30.39	i	3	Defective prophage *rac*
radC	252	82	82.31	nc	3	Sensitivity to radiation
rafA				ns		α-Galactosidase
rafR	1,011			nm	3	Raffinose repressor
rbsA	1,506	84	85.03	d	1	Ribose; D-ribose high-affinity transport protein; membrane-associated protein
rbsB	891	84	85.08	d	2	Ribose; D-ribose periplasmic binding protein precursor
rbsC	966	84	85.06	d	1	Ribose; D-ribose high affinity transport system; membrane-associated protein
rbsD	420	84	85.02	d	1	Ribose; D-ribose high affinity transport system; membrane-associated protein
rbsK	930	84	85.10	d	1	Ribose; ribokinase (EC 2.7.1.15)
rcsA	624	43	43.31	nc	3	Positive regulatory gene for capsule synthesis
rcsB	651	48	92.13	nc	1	Positive regulatory gene for capsule synthesis
rcsC	2,802	48	92.07	nc	1	Negative regulatory gene for capsule synthesis
recA	1,059	58	60.07	nc	2	Recombinaison; general recombination, repair of radiation damage, and induction of phage lambda
recB	3,540	61	62.86	d	1	Recombinaison; exonuclease V (EC 3.1.11.5) subunit; DNA recombination; DNA repair
recC	3,366	61	63.00	d	1	Recombinaison; exonuclease V (EC 3.1.11.5) subunit; DNA recombination; DNA repair
recD	1,824	61	62.82	d	1	Recombinaison; exonuclease V (EC 3.1.11.5) alpha subunit; recombination and DNA repair
recE		30	30.38	ns		Recombinaison; locus of *rac* prophage; recombinaison and DNA repair; exonuclease VIII
recF	1,071	83	83.78	d	1	Recombinaison; recombination and repair of radiation damage
recG	2,082		82.60	i	1	Recombinaison; DNA recombinaison
recJ	1,737	62	64.65	d	1	Recombinaison; recombinaison and DNA repair
recN	1,701	57	58.42	i	1	Recombinaison; DNA repair; DNA recombination
recO	726	56	57.31	d	1	Recombinaison; conjugational recombination and DNA repair
recQ	1,833	86	86.54	d	1	Recombinaison; conjugational recombination and DNA repair
recR	606		10.73	d	1	Recombinaison; DNA recombination
relA	2,235	60	62.00	d	1	Relaxed; regulation of RNA synthesis; stringent factor; ATP:GTP 3′-pyrophosphotransferase
relB	237	35	35.22	d	2	Relaxed; regulation of RNA synthesis
relE	285	35	35.21	d	3	Relaxed; regulation of RNA synthesis; function unknown
relF	153	35	35.21	d	3	Relaxed; regulation of RNA synthesis; function unknown
rem	249		35.20	d	3	Function unknown
rep	1,911	85	85.58	d	1	Rep helicase; a single-stranded DNA dependant ATPase
Rev*893*	1,134			nm	1	Biocyclomycin resistance
rfa-2	960		81.97	i	3	Rough; function unknown
rfaD	933	81	81.93	i	2	Rough; ADP-L-glycero-D-mannoheptose-6-epimerase
rfaG	1,125	81	82.24	i	3	Rough; lipopolysaccharide core biosynthesis; glucosyltransferase I
rfaP	798	81	82.26	i	3	Rough; lipopolysaccharide core biosynthesis; phosphorylation of core heptose
rfaQ	969	81	82.22	i	3	Rough; lipopolysaccharide core biosynthesis
rfe	774	85	85.76	d	1	Rough; Involved in synthesis of enterobacterial common antigen and O antigen
rhaR	936	88	88.47	i	1	Rhamnose; positive activator of genes for L-rhamnose utilization
rhaS	834	88	88.49	i	1	Rhamnose; positive activator of genes for L-rhamnose utilization
rhlB	1,266		85.68	i	1	RNA helicase-like protein
rho	1,260	85	85.72	d	2	Transcription termination factor Rho; polarity suppressor
rhsA	4,134	81	81.24	i	1	Repetitive sequence responsible for duplications within chromosome
rhsB		77	78.20	ns		Repetitive sequence responsible for duplications within chromosome
rhsC		16	15.85	ns		Repetitive sequence responsible for duplications within chromosome
rhsD	4,281	12	11.36	d	1	Repetitive sequence responsible for duplications within chromosome
rhsE			31.33	ns		Repetitive sequence responsible for duplications within chromosome
ribA		28	28.71	ns		Riboflavin; GTP cyclohydrolase II
rimI	483	99	99.40	nc	1	Ribosomal modification; modification of 30S ribosomal subunit protein S18; acetylation of N-terminal alanine
rimJ	582	32	24.21	d	1	Ribosomal modification; modification of 30S ribosomal subunit protein S5; acetylation of N-terminal alanine
rimK	876		19.73	nc	1	Ribosomal modification; ribosomal protein S6 modification
rimL	537	33	32.86	i	1	Ribosomal modification; modification of 30S ribosomal subunit protein L7; acetylation of N-terminal serine
rlpA	1,089	15	14.41	i	1	A minor lipoprotein
rlpB	582	15	14.59	i	1	A minor lipoprotein
rna	807	14	13.84	i	1	RNase; RNase I
rnc	681	55	57.35	d	1	RNase; RNase III
rnd	1,128	40	40.41	d	1	RNase; RNase D

Continued on following page

TABLE 1—*Continued*

Name	Length (bp)	Map position		*oriC*[c]	Class[d]	Description
		M.G[a]	M.P[b]			
rnh	465	5	5.25	i	1	RNase; RNase H
rnpA	360	83	83.87	i	1	RNase; RNase P, protein component
rob	870		99.84	i	1	Right origin-binding protein
rodA	1,113	15	14.44	i	1	Cell shape; sensitivity to radiation and drugs
rplA	702	90	90.23	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L1
rplB	819	73	74.64	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L2
rplC	627	73	74.67	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L3
rplD	603	73	74.66	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L4
rplE	537	73	74.56	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L5
rplF	531	73	74.53	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L6
rplI	447	96	95.48	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L9
rplJ	495	90	90.25	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L10
rplK	426	90	90.22	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L11
rplL	363	90	90.27	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L7/L12
rplM	426	70	73.02	nc	2	Ribosomal protein; large; 50S ribosomal subunit protein L13
rplN	369	73	74.58	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L14
rplO	432	73	74.50	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L15
rplP	408	73	74.60	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L16
rplQ	381	73	74.41	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L17
rplR	351	73	74.52	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L18
rplS	345	57	58.26	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L19
rplT	357	38	38.48	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L20
rplV	330	73	74.62	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L22
rplW	300	73	74.65	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L23
rplX	312	73	74.57	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L24
rpmB	237	82	82.29	nc	2	Ribosomal protein; large; 50S ribosomal subunit protein L28
rpmC	189	73	74.60	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L29
rpmD	177	73	74.51	d	2	Ribosomal protein; large; 50S ribosomal subunit protein L30
rpmF	174	24	24.67	nc	2	Ribosomal protein; large; 50S ribosomal subunit protein L32
rpmG	168	82	82.30	nc	2	Ribosomal protein; large; 50S ribosomal subunit protein L33
rpmH	141	83	83.87	i	2	Ribosomal protein; large; 50S ribosomal subunit protein L34
rpoA	987	73	74.41	d	2	RNA polymerase; RNA polymerase (EC 2.7.7.6), alpha subunit
rpoB	4,026	90	90.28	d	2	RNA polymerase; RNA polymerase (EC 2.7.7.6), beta subunit
rpoC	4,221	90	90.37	d	2	RNA polymerase; RNA polymerase (EC 2.7.7.6), beta′ subunit
rpoD	1,839	67	69.26	i	2	RNA polymerase; RNA polymerase (EC 2.7.7.6), sigma 70 subunit
rpoN	1,434	70	72.90	nc	1	RNA polymerase; RNA polymerase (EC 2.7.7.6), sigma 60 subunit; enhancer factor σ^{54}
rpoZ	276		82.52	i	2	RNA polymerase; omega protein
rpsA	1,668	21	20.67	d	2	Ribosomal protein, small; 30S ribosomal subunit protein S1
rpsB	723	4	4.28	d	2	Ribosomal protein, small; 30S ribosomal subunit protein S2
rpsC	699	73	74.61	d	2	Ribosomal protein, small; 30S ribosomal subunit protein S3
rpsD	618	73	74.44	d	2	Ribosomal protein, small; 30S ribosomal subunit protein S4
rpsE	501	73	74.51	d	2	Ribosomal protein, small; 30S ribosomal subunit protein S5
rpsF	393	95	95.46	d	2	Ribosomal protein, small; 30S ribosomal subunit protein S6
rpsG		73	75.12	ns		Ribosomal protein, small; 30S ribosomal subunit protein S7
rpsH	390	73	74.54	d	2	Ribosomal protein, small; 30S ribosomal subunit protein S8
rpsI	390	70	73.03	nc	2	Ribosomal protein, small; 30S ribosomal subunit protein S9
rpsJ	309	73	74.69	d	2	Ribosomal protein, small; 30S ribosomal subunit protein S10
rpsK	387	73	74.45	d	2	Ribosomal protein, small; 30S ribosomal subunit protein S11
rpsL	372	73	75.12	d	2	Ribosomal protein, small; 30S ribosomal subunit protein S12
rpsM	354	73	74.46	d	1	Ribosomal protein, small; 30S ribosomal subunit protein S13
rpsN	297	73	74.55	d	2	Ribosomal protein, small; 30S ribosomal subunit protein S14
rpsO	270	69	71.57	d	2	Ribosomal protein, small; 30S ribosomal subunit protein S15
rpsP	246	57	58.30	d	2	Ribosomal protein, small; 30S ribosomal subunit protein S16
rpsQ	252	73	74.59	d	2	Ribosomal protein, small; 30S ribosomal subunit protein S17
rpsR	225	96	95.48	d	2	Ribosomal protein, small; 30S ribosomal subunit protein S18
rpsS	276	73	74.63	d	2	Ribosomal protein, small; 30S ribosomal subunit protein S19
rpsT	261	0	0.46	i	2	Ribosomal protein, small; 30S ribosomal subunit protein S20
rpsU	213	67	69.22	i	2	Ribosomal protein, small; 30S ribosomal subunit protein S21
rrfX	558		4.46	d	2	Ribosome-releasing factor
rsgA	498		42.35	nc	2	Function unknown; homology to human ferritin subunit
ruvA	612	41	41.60	nc	1	Filament formation and sensitivity to UV radiation
ruvB	1,011	41	41.61	nc	1	Filament formation and sensitivity to UV radiation
ruvC	522			nm	1	DNA repair; DNA recombination; resolvase of Holliday junction intermediates
sbcB	1,401	44	44.35	i	1	Exonuclease I; suppression of *recB*, *recC* mutations
sbcC	3,144	9	9.02	i	1	Suppression of *recB*, *recC* mutations
sbmA	1,221	9	8.28	nc	1	Sensitivity to microcin B17

Continued on following page

TABLE 1—*Continued*

Name	Length (bp)	Map position		*oriC*[c]	Class[d]	Description
		M.G[a]	M.P[b]			
sbp	987	89	88.75	d	1	Periplasmic sulfate-binding protein
sdaA	1,347			nm	1	L-Serine deaminase
sdhA	1,767	16	16.29	d	2	Succinate dehydrogenase (EC 1.3.99.1), flavoprotein subunit
sdhB	717	16	16.32	d	1	Succinate dehydrogenase (EC 1.3.99.1), iron sulfur protein
sdhC	390	16	16.27	d	3	Succinate dehydrogenase (EC 1.3.99.1); cytochrome b_{556}
sdhD	348	16	16.28	d	1	Succinate dehydrogenase (EC 1.3.99.1), hydrophobic subunit
secA	2,706	2	2.35	d	2	Secretion of envelope proteins
secB	468	81	81.76	nc	2	Cytoplasmic export protein
secD	1,845	9	9.32	d	2	Membrane proteins involved in protein export
secE	384		90.20	d	1	Protein export
secF	969		9.36	d	2	Membrane proteins involved in protein export
secY	1,329	73	74.47	d	1	Protein export; membrane protein
selA	1,392	81	81.04	nc	1	Selenium; selenium biosynthesis; selenocysteine synthase
selB	1,842	81	81.12	i	1	Selenium; selenium metabolism; biosynthesis or incorporation of selenocystein
selD	1,041		39.55	nc	1	Selenium; selenium metabolism protein
serA	1,233	63	65.06	d	1	Serine; D-3-phosphoglycerate dehydrogenase (EC 1.1.1.95)
serB	966	100	99.66	d	1	Serine; phosphoserine phosphatase (EC 3.1.3.3)
serS	1,290	20	20.19	d	2	Serine; seryl-tRNA synthetase (EC 6.1.1.11)
sfaA	543			nm	3	Fimbriae; S-fimbrial protein
sfs-1	705		3.46	d	1	Sugar fermentation stimulation protein 1
skp	486	4	4.50	d	2	Histonelike protein HLP-I (BH1); DNA-binding nucleoid-associated protein
slt	1,938		99.77	d	1	Soluble lytic transglycosylase
sms	1,383		99.68	d	1	Function unknown
sodA	618	88	88.58	nc	2	Superoxide dismutase, manganese
sodB	582	36	37.15	i	2	Superoxide dismutase, iron
sohB	648		28.51	d	1	Multicopy suppressor of the *htrA* (*degP*) null phenotype; transmembrane protein
soxR	465		92.32	nc	1	Regulatory gene for superoxide stress response
soxS	324		92.31	nc	1	Regulatory gene for superoxide stress response
speA	1,977	64	65.65	d	2	Spermidine; arginine decarboxylase (EC 4.1.1.19)
speB	921	64	65.63	d	2	Spermidine; agmatinase (EC 3.5.3.11)
speC	2,196	64	66.14	d	1	Spermidine; ornithine decarboxylase (EC 4.1.1.17)
speD	795	3	2.95	nc	1	Spermidine; *S*-adenosylmethionine decarboxylase (EC 4.1.1.50)
speE	867	3	2.93	nc	1	Spermidine; spermidine synthase (putrescine aminopropyltransferase) (EC 2.5.1.16)
speF	2,199		15.53	i	1	Spermidine; ornithine decarboxylase?
spoR	624		82.50	i	1	Guanosine; 5′-guanylate kinase; GMP kinase (EC 2.7.4.8)
spoT	2,109	82	82.53	i	1	Guanosine; guanosine 3′,5′-bis(diphosphate) 3′-pyrophosphatase
spoU			82.59	ns		Guanosine; function unknown
sppA	1,857	39	39.59	i	1	Protease IV, a signal peptide peptidase
srmB	1,335			nm	1	eIF-4A like protein
ssb	537	92	92.25	d	2	Single-strand DNA-binding protein
sspB	498	70	72.72	nc	1	Stringent starvation protein B
sspG	636	70	72.70	nc	2	Stringent starvation protein
sucA	2,802	16	16.34	d	2	Succinate; α-ketoglutarate dehydrogenase, decarboxylase component
sucB	1,218	16	16.40	d	2	Succinate; α-ketoglutarate dehydrogenase, dihydrolipoyltranssuccinase component
sucC	1,167	16	16.44	d	2	Succinate; succinyl-CoA synthetase (EC 6.2.1.5), beta subunit
sucD	870	16	16.46	d	2	Succinate; succinyl-CoA synthetase (EC 6.2.1.5), alpha subunit
sufI			68.17	ns		Periplasmic protein; suppresses *ftsI* mutation
suhB	804		56.48	i	2	Extragenic suppressor
sulA		22	21.95	ns		Suppressor of *lon*
tag	564	72	80.20	nc	1	3-Methyl-adenine DNA glycosylase I, constitutive
tap	1,605	42	41.97	d	1	Methyl-accepting chemotaxis protein IV
tar	1,659	42	42.01	d	1	Methyl-accepting chemotaxis protein II
tau	930		36.06	i	1	DNA replication terminus site-binding protein
tdcA	936	68	70.62	d	3	Threonine dehydratase (EC 4.2.1.16)
tdcB	987	68	70.60	d	1	Threonine dehydratase (EC 4.2.1.16)
tdcC	1,293	68	70.57	d	1	Threonine dehydratase (EC 4.2.1.16)
tdcR	297	68	70.64	i	3	Threonine dehydratase (EC 4.2.1.16)
tdh	1,023	81	81.85	d	1	Threonine dehydrogenase (EC 1.1.1.103)
tdk	618	27	27.72	d	3	Thymidine kinase (EC 2.7.1.75)
tesB	861	10	10.31	i	1	Thioesterase II
tgt	1,128	9	9.29	d	1	tRNA guanine transglycosylase
tgy	102		27.61	i	3	Protamine-like protein
thdF	1,320		83.97	nc	3	Thiophene degradation; thiophene and furan oxidation
thrA	2,463	0	0.01	d	1	Threonine; aspartokinase I (EC 2.3.2.4), homoserine dehydrogenase I (EC 1.1.1.3)

Continued on following page

TABLE 1—*Continued*

Name	Length (bp)	Map position		*oriC*[c]	Class[d]	Description
		M.G[a]	M.P[b]			
thrB	930	0	0.06	d	1	Threonine; homoserine kinase (EC 2.7.1.39)
thrC	1,287	0	0.08	d	1	Threonine; threonine synthase (EC 4.2.99.2)
thrS	1,929	38	38.51	d	2	Threonine; threonyl-tRNA synthetase (EC 6.1.1.3)
thyA	792	61	63.11	d	1	Thymine; thymidylate synthetase (EC 2.1.1.45)
tig	1,299		9.92	nc	2	Trigger factor
tnaA	1,416	84	84.00	nc	1	Tryptophanase (EC 4.1.99.1)
tnaB	1,248		84.01	nc	1	Low-affinity tryptophan permease
tnaC	75		84.00	nc	3	Tryptophanase (EC 4.1.99.1)
tnpA	1,281			nm	1	Transposase IS*91*
tolA	1,266	17	16.76	d	3	Tolerance; tolerance to group A colicins and single-stranded filamentous DNA phages
tolB	1,296	17	16.79	d	2	Tolerance; tolerance to colicins E2, E3, A, and K
tolC	1,485	66	68.51	i	1	Tolerance; specific tolerance to colicin E1; expression of outer membrane proteins
tolQ	288	17	16.73	d	3	Tolerance; tolerance to group A colicins and single-stranded filamentous DNA phages
tolR	681	17	16.73	d	1	Tolerance; tolerance to group A colicins and single-stranded filamentous DNA phages
tonB	735	28	28.12	nc	1	T1; uptake of chelated iron and cyanocobalimin; sensitivity to phages T1 and φ80 and colicins
topA	2,595	28	28.55	d	1	Topoisomerase; DNA topoisomerase I, omega protein
topB	1,962		39.52	nc	1	Topoisomerase; topoisomerase III
tpi	765	89	88.79	i	2	Triose phosphate isomerase (EC 5.3.1.1)
treA	1,695	26	26.67	i	1	Trehalose; trehalase, periplasmic
trg	1,608	31	32.00	i	3	Methyl-accepting chemotaxis protein III
trkA	1,377	72	74.27	nc	1	Transport of potassium; protein of the constitutive K^+ transport system
trkG	1,458		30.55	nc	3	Protein involved in potassium uptake via the Trk system
trmA	1,101	90	89.87	i	1	tRNA methyltransferase; tRNA (uracil-5)-methyltransferase (EC 2.1.1.35)
trmD	765	57	58.27	d	1	tRNA methyltransferase; tRNA (guanine-7)-methyltransferase (EC 2.1.1.31)
trpA	807	28	28.24	i	1	Tryptophan; tryptophan synthase (EC 4.2.1.20), A protein
trpB	1,194	28	28.25	i	1	Tryptophan; tryptophan synthase (EC 4.2.1.20), B protein
trpC	1,359	28	28.28	i	1	Tryptophan; *N*-(5-phosphoribosyl)anthranilate isomerase indole-3-glycerol phosphate synthetase
trpD	1,596	28	28.31	i	1	Tryptophan; glutamine aminotransferase-phosphoribosyl anthranilate transferase
trpE	1,563	28	28.34	i	1	Tryptophan; anthranilate synthase (EC 4.1.3.27)
trpR	327	100	99.81	d	1	Tryptophan; regulation of *trp* operon and *aroH*; *trp* aporepresseur
trpS	1,002	74	75.95	nc	1	Tryptophan; tryptophanyl-tRNA synthetase (EC 6.1.1.2)
trxA	330	86	85.71	d	2	Thioredoxin; thioredoxin deficiency
trxB	966	21	20.07	nc	1	Thioredoxin; thioredoxin reductase
tsf	849	4	4.30	d	2	Protein chain elongation factor EF-Ts
tsr	1,608	99	99.07	i	1	Methyl-accepting chemotaxis protein I
tsx	885	9	9.39	d	2	Outer membrane protein; nucleoside uptake; receptor for phage T6 and colicin K
tufA	1,185	74	75.05	d	2	Protein chain elongation factor EF-Tu (duplicate gene)
tufB		90	90.17	ns	2	Protein chain elongation factor EF-Tu (duplicate gene)
tyrA	1,119	57	58.20	nc	1	Tyrosine; chorismate mutase-T (EC 5.4.99.5); prephenate dehydrogenase (EC 1.3.1.12)
tyrB	1,194	92	92.12	d	1	Tyrosine; tyrosine aminotransferase (EC 2.6.1.5); tyrosine repressible
tyrR	1,419	29	28.84	nc	1	Tyrosine; regulation of *aroFG* and *tyrA* and aromatic amino acid transport systems
tyrS	1,275	36	36.46	d	2	Tyrosine; tyrosyl-tRNA synthetase (EC 6.1.1.1)
ubiA	873	92	91.79	d	1	Ubiquinone; enzymatic 3-octaprenyl-4-hydroxybenzoate synthesis; 4-hydroxybenzoate octaprenyl transferase
ubiC	498	92	91.78	d	1	Ubiquinone; chorismate lyase; enzymatic chorismate→*p*-hydroxybenzoate + pyruvate
ubiH	1,179	63	64.34	nc	1	Ubiquinone; 2-octaprenyl-6-methoxyphenol→2-octaprenyl-6-methoxy-1,4-benzoquinone
udp	759	86	86.78	d	2	Uridine phosphorylase (EC 2.4.2.3)
ugpA	885	76	77.62	d	1	*sn*-Glycerol 3-phosphate transport system
ugpB	1,314	76	77.64	d	1	Binding protein of *sn*-glycerol 3-phosphate transport system
ugpC	1,068	76	77.58	d	1	*sn*-Glycerol 3-phosphate transport system
ugpE	843	76	77.60	d	1	*sn*-Glycerol 3-phosphate transport system; membrane protein
ugpQ	741		77.56	d	1	Glycerophosphoryl diester phosphodiesterase
uhpA	591	82	83.13	d	1	Hexose phosphate transport protein; positive activator of *uhpT* transcription
uhpB	1,557	82	83.09	d	1	Hexose phosphate transport protein; regulatory gene
uhpC	660	82	83.07	d	1	Hexose phosphate transport protein; regulatory gene
uhpT	1,392	82	83.03	d	1	Hexose phosphate transport protein; transport protein
uidA	1,809	36	36.30	d	1	β-D-Glucuronidase (EC 3.2.1.31)

Continued on following page

TABLE 1—*Continued*

Name	Length (bp)	Map position		*oriC*[c]	Class[d]	Description
		M_G[a]	M_P[b]			
umuC	1,269	26	26.38	d	1	Induction of mutations by UV; error-prone repair; sensitive to UV
umuD	420	26	26.38	d	1	UV mutagenesis; Induction of mutation by UV; error-prone repair
uncA	1,539	84	84.70	d	2	ATP; membrane-bound ATP synthase (EC 3.6.1.3); F1 sector, alpha subunit
uncB	813	84	84.76	d	1	ATP; membrane-bound ATP synthase (EC 3.6.1.3); F0 sector, subunit a
uncC	417	84	84.64	d	2	ATP; membrane-bound ATP synthase (EC 3.6.1.3); F1 sector, epsilon subunit
uncD	1,380	84	84.65	d	2	ATP; membrane-bound ATP synthase (EC 3.6.1.3); F1 sector, beta subunit
uncE	237	84	84.75	d	2	ATP; membrane-bound ATP synthase (EC 3.6.1.3); F0 sector, subunit c
uncF	468	84	84.74	d	2	ATP; membrane-bound ATP synthase (EC 3.6.1.3); F0 sector, subunit b
uncG	861	84	84.68	d	2	ATP; membrane-bound ATP synthase (EC 3.6.1.3); F1 sector, gamma subunit
uncH	531	84	84.73	d	1	ATP; membrane-bound ATP synthase (EC 3.6.1.3); F1 sector, delta subunit
uncI	390	84	84.77	d	3	ATP; membrane-bound ATP synthase (EC 3.6.1.3)
ung	690	56	57.65	nc	1	Uracil-DNA-glycosylase
upp	627	54	55.48	nc	2	Uracil phosphoribosyltransferase (EC 2.4.2.9)
usg	1,011		51.89	d	1	Function unknown
ushA	1,650	11	11.10	nc	1	UDP-glucose hydrolase (F'-nucleotidase)
uvrA	2,823	92	92.19	i	1	UV; repair of UV damage to DNA; exision nuclease
uvrB	2,019	18	17.53	d	1	UV; repair of UV damage to DNA; exision nuclease
uvrC	1,764	42	42.47	d	1	UV; repair of UV damage to DNA; exision nuclease
uvrD	2,160	86	86.37	d	1	UV; repair of UV damage to DNA; DNA-dependent ATPase I and DNA helicase II
uxaB		52	35.24	ns		Altronate oxidoreductase (EC 1.1.1.58)
uxuA		98	98.50	nc		Mannonate hydrolase (EC 4.2.1.8)
valS	2,853	97	96.55	i	2	Valine; valyl-tRNA synthetase (EC 6.1.1.9)
visA	963		10.80	d	1	Visible-light sensitivity; function unknown
visC	1,203		64.36	nc	1	Visible-light sensitivity; function unknown
witA	1,299		57.83	nc	3	α-Ketoglutarate transporter
xerB	1,509		96.63	i	1	Aminopeptidase A/I
xerC	894		86.34	d	1	Lambda integrase
xprA	708		64.69	d	1	Function unknown
xprB	897		64.70	d	1	Function unknown
xseA	1,371	54	55.87	nc	1	Exonuclease VII, large subunit
xthA	804	38	39.11	nc	1	Exonuclease III
xylA	1,335	80	80.57	d	1	Xylose; D-xylose isomerase (EC 5.3.1.5)
xylB	1,455	80	80.54	d	1	Xylose; xylulose kinase (EC 2.7.1.17)
xylE	1,476	91	91.54	i	1	Xylose; xylose-proton symport
xylUP	183			nm	3	Xylose; D-xylose uptake protein
zwf	1,473	41	41.30	nc	1	Zwischenfement; glucose-6-phosphate dehydrogenase (EC 1.1.1.49)

[a] M_G, experimental genetic map position (1).
[b] M_P, Map position in minutes, calculated from the genomic address of the gene transcription start in the *E. coli* restriction map (16).
[c] *oriC* direction of genes transcription with respect to the replication forks. d represents clockwise direction, and i represents counterclockwise direction. We used nm when the map localization is not known, nc when the Kohara restriction map is not corrected for the corresponding position of the gene (19), and ns when a gene is not entirely sequenced.
[d] The number obtained after statistical analysis defining *E. coli* genes classes according to their codon usage (see text and reference 21).
[e] CoA, coenzyme A.

GRAPHIC REPRESENTATION OF DATA STORED IN COLIBRI

At this stage, we have described the two basic steps necessary for the building up of the data base dedicated to the analysis of the *E. coli* genome, its logical structure, and the environment for data consultation. It should be emphasized here that such a data organization has allowed us to realize "natural" connections between them (Fig. 3). It is thus possible, without any complementary software development, to extract new information, which will be presented in a graphic form below.

Sequencing Density

The first question about nonredundant DNA sequences of the *E. coli* genome is that of the number of known DNA sequences and their distribution along the chromosome. To illustrate this feature of Colibri, we have represented in Fig. 7 the "sequencing density" of the *E. coli* genome, i.e., the distribution of known sequences calculated as the number of sequenced nucleotides per 50 kbp of physically mapped chromosome length, limited to the knowledge accumulated up to July 1992 in the EMBL data library. The black region calculated from the physical map coordinates and from the length of each contig localized in the corrected Kohara map (19) represents the sequenced area. This represents more than 40% of the total genome. This region now reaches the outer circle (at 0 to 2.4 min and at 84.5 to 86.5), since two contigs of about 100 kbp long have just been entirely sequenced (7, 31). From the figure it appears that the known sequences are unevenly distributed. Some gaps are localized in well-sequenced regions (for example, at 6 and 80 min), suggesting that all the *E. coli* genetic information is not accessible by only mutant phenotypes studies.

Genetic Map and Physical Map

We have previously shown that the determination of the genomic address of a contig on the *E. coli* physical map

E. coli genes : 1276

selected : 13

GENE	Length (pb)	M_G	M_P	Pos kb	oriC	Description	Clas
dnaK	1917	0	0,27	12,93	d	DNA biosynthesis; heat shock protein	2
dnaJ	1131	0	0,32	14,94	d	DNA biosynthesis	2
polB	2307	2	1,44	65,85	i	DNA polymerase II	1
dnaE	3438	4	4,6	217,34	d	DNA polymerase IIII, alpha subunit	1
mutD	729	5	5,26	248,15	d	Mutator activity; DNA polymerase III holoenzyme, epsilon subunit	1
dnaZX	1932	11	10,68	504,02	d	DNA polymerase IIII, gamma subunit; DNA elongation factor III	1
dnaG	1743	67	69,22	3267,27	d	DNA biosynthesis; primase	1
dnaN	1101	83	83,82	3956,41	i	DNA polymerase IIII, beta subunit	1
dnaA	1404	83	83,85	3957,82	i	DNA biosynthesis; initiation	1
polA	2787	87	87,44	4126,95	d	DNA polymerase I	1
dnaB	1416	92	-	-	-	DNA biosynthesis; chain elongation	1
dnaC	738	99	99,15	4680,84	i	DNA biosynthesis; initiation and chain elongation	1
dnaT	540	99	99,16	4682,43	i	DNA biosynthesis; primasomal protein I	1

Operations on the selection		Find	Consultation of	
Sort	Sub-selection	general information	>> CONTIGS	>> Scan FastP
Print	Show all cards	Function	>> EMBL	Return

FIG. 6. Layout for *E. coli* gene information in list form. This figure presents a selection of records of the file [GENES]; it was obtained after searching for genes linked to the keyword "DNA biosynthesis." The data presented in this layout indicate for each selected gene its name and length (pb), its genetic map (M_G) and physical map (M_P) positions, its genomic address (Pos kb) and orientation of transcription with respect to the replication forks (oriC; d for clockwise and i for counterclockwise), its description, and finally a number corresponding to the class to which it belong according to its codon usage (Class column; see text and Table 1).

allows calculation of an exact map position (in minutes) for the corresponding genes. It was interesting to compare the experimental map positions with these physical map positions. In our data base, a total of 700 genes are sufficiently information rich for the fields corresponding to the published genetic map location (M_G) and the physical map position (M_P). We find that, except for *ponB*, *hns*, and *ftsQ* localizations, the conformity between the physical map and genetic map is very good. We find, however, that genetic map positions are underestimated in the interval from 50 to 80 min. This effect has been previously noticed by Rudd et al. (26). A biological significance of these variations could come from the rate of transfer of different regions of the chromosome during Hfr crosses. It is known that the portion of the chromosome that is the slowest to be transferred is the region that contains most of the actively transcribed ribosomal protein and RNA genes (4), localized mainly arround 70 min.

Gene Transcription versus Replication

Taking into account information calculated in our data base, one can also define the direction of gene transcription with respect to the replication forks. Genes on the *E. coli* chromosome are transcribed in either a clockwise or counterclockwise direction relative to the standard Kohara map (16). Knowing the map position and the direction of transcription relative to flanking markers, the direction of transcription for a given gene relative to the origin (*oriC*) at 84 min can be determined. In Fig. 8 we have represented the length distribution of genes transcribed away from the origin (black region) and of genes transcribed toward the origin (hatched region). As shown in Fig. 8, the orientation preference is related to transcription itself, especially arround the origin (84 min) but also at 75 min and near 2 min. The two peaks corresponding to genes transcribed in a counterclockwise direction with respect to the replication forks are not very significant, since they are in sequenced regions also containing genes transcribed in a clockwise direction (Fig. 8). With regard to the terminus region (30 min), results are

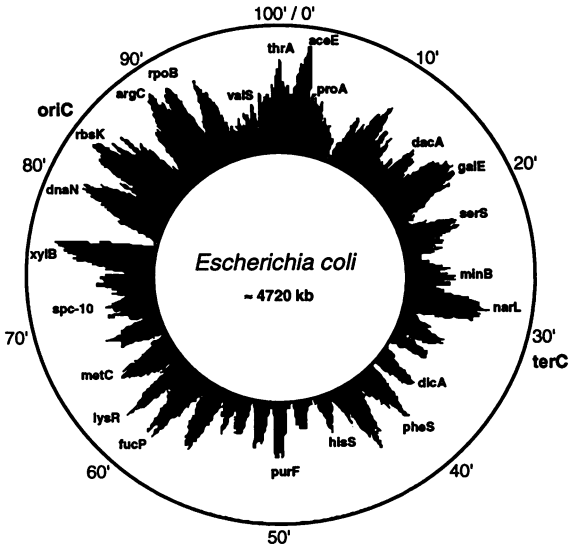

FIG. 7. Distribution of known *E. coli* DNA fragments along the chromosome. The black region, i.e., the sequenced area, is calculated from the physical map coordinate and the length of each contig localized on the corrected Kohara map (20, 22). For each position along the chromosome, the number of sequenced nucleotides per 50 kbp is plotted. Some genetic markers are included for comparison with the genetic map (1).

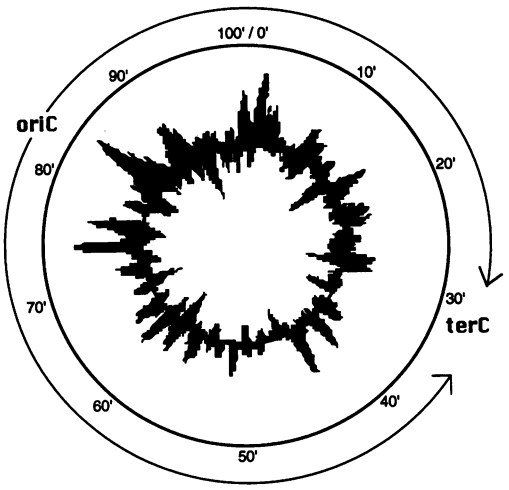

FIG. 8. Gene transcription versus replication. This figure represents the length distribution of genes transcribed in a clockwise direction with respect to the replication forks (hatched region), and the length distribution of genes transcribed in a counterclockwise direction (black region). For each position along the chromosome, the number of nucleotides contained in genes per 50 kbp is plotted. This density is calculated only for genes located at known positions in the *E. coli* chromosome.

not significant because the exact point of the end of the replication is not yet known precisely enough. In fact, interrogation of our data base by using the biological gene function criterion has shown that there is more likely to be a correspondence between the direction of gene transcription and the degree of gene expression, i.e., the degree of transcriptional activity. For example, it was found that more actively transcribed genes, such as the genes (i.e., genes from class 2) involved in the protein synthesis machinery, are in most cases transcribed away from *oriC*.

COLIBRI AS A MODEL FOR MICROBIAL GENOME DATA BASES

Colibri is a data base that permits recovery of self-consistent, nonredundant DNA sequences of the *E. coli* genome. The corresponding information is extracted from existing data libraries in such a way that data management can be easily performed by using a set of ad hoc procedures that have been implemented by keeping in mind the constraint that the general user does not have to identify the underlying structure of the base. The data structure has been designed to permit the aggregation and recovery of the generally fuzzy knowledge associated with the molecular genetics of *E. coli*. Appropriate links created between the different objects present in the data base permit rapid and direct access to the variety of biological information. The present state of Colibri makes it much more complete than the data base described by Kunisawa et al. (18), which contained only 20% of the total DNA content of the *E. coli* genome. This has already permitted a thorough exploration of the general properties of the genome, as now summarized.

A first consequence of the magnitude of the data present in Colibri is that the overall organization of the *E. coli* genome can be analyzed. In particular, it has been possible to correlate the physical map established for one laboratory strain (W3110) by Kohara et al. with the vast but extremely scattered set of sequences produced by the many laboratories interested in *E. coli* genetics. This permitted us to show that, with the exception of the 52-min region and the borders of the chromosomal inversion present in W3110 in contrast to most strains of *E. coli*, the physical map of Kohara et al. is extremely accurate (19). This led us to demonstrate that polymorphism among laboratory K-12 strains of *E. coli* is extremely low (of the order of 10^{-3} base change per base) and of the same order of magnitude as the usual error rate generated during sequencing or recording sequences in data libraries (11). This also allowed for the correction of the physical map and led us to discover that two restriction enzymes (*Eco*RV and *Pvu*II) were context sensitive and that the distribution of *Eco*RV, *Hin*dIII, and *Pst*I sites differed from the expected random Poisson distribution (22). This latter observation has been substantiated by several thorough independent analyses (6, 13). A further statistical analysis, in which the constraints of the codon usage in coding sequences were used as a reference, has also demonstrated the interest for a thorough statistical study of the genome. We can expect to discover interesting biological properties from such analysis in the near future (21). As a case in point, it can be demonstrated, by using the data base, that Chi sites (GCTGGTGG) are distributed in a highly nonrandom fashion; i.e., they usually have the same orientation as that of the replicating fork. This indicates that, from analysis of biases in the sequence one should be able to recognize replication orientation (Table 2).

TABLE 2. Occurrence of Chi sites in coding sequences compared with replication and transcription orientation

Chi site	Occurrence in[a]:						Total
	Td and Rd	Td and Ri	Ti and Rd	Ti and Ri	Td and R?	Ti and R?	
GCTGGTGG	95	13	20	9	49	0	186
CCACCAGC	6	18	28	76	13	7	148

[a] Abbreviations: T, transcription; R, replication; d, direct; i, indirect; R?, when direction of genes transcription with respect to the replication forks is not known.

As a model study in data processing, the present work also has a methodological interest. Indeed, it provides us with a paradigm for the building up of evolving data bases for implementing genomic information, at least for procaryotes. This is presently used for *Bacillus subtilis*, whose genome is being sequenced by a team of European and Japanese groups. In this context, however, the appropriate interfacing with the raw data obtained immediately downstream from the sequencing gels has yet to be defined. The specific relational data base management system that was used, 4D, permitted us to define an organization of the data that allowed rational analysis of sequence data. After each run of analytical procedures progressively added to the core of methods specific to Colibri, the knowledge that has already been implemented in the data base is modified and improved. Therefore, although it is necessary to control the consistency and integrity of new data added to the base, automatic procedures modify certain fields as a function of the actual use of the data base, as well as during updating. This requires the writing and management of multiple internal procedures. It should be noted that adding new information can be extremely tedious. Therefore, whereas a relational data base management system appears to be able to satisfy most of the requisites induced by the analysis of whole genomes, it may lack some of the flexibility needed for the proper functioning of an "evolving" data base. As a conse-

quence, it seems interesting to consider object-oriented data base management system for the future. There are already several examples. An object-oriented knowledge base, ColiGene (24), has been developed by using the knowledge base management system SHIRKA (25); it aims to study the expression of *E. coli* genes, taking into account the expertise of molecular geneticists. Another object-oriented system, ACeDB, has been specifically constructed for the management of sequences and knowledge induced by the program of *Caenorhabditis elegans* genome sequencing (29). This latter data base is now used by all laboratories involved in this program, and it has been chosen as a model for the development of other specialized data bases (those for *Drosophila* species, mouse, etc.). Finally, the data records and the procedures created by Rudd et al. have been recently implemented in an object-oriented data base (28). The corresponding prototype is, however, too new to permit evaluation of the advantages of object-oriented with respect to relational data base management systems. The experience acquired during the building up of Colibri leads us naturally to be interested in these latter approaches. As a first step we shall presently use SHIRKA, aiming in particular at evaluating problems posed to the persistence and integrity of data, for heavy-duty data bases, such as those which should follow Colibri.

The Colibri data base, embedded in a runtime version of 4D, can be obtained through anonymous ftp. Users connected to the Internet network can type 'ftp radium.jussieu.fr' (or, in case of difficulties, 'ftp 134.157.56.1'), enter 'anonymous' as the user identification and any word as password to access to the repository root directory. Users without ftp experience should then issue a 'get radium.readme' command and should carefully read this file on their own site. The Colibri repository directory (/pub/colibri) also contains a 'colibri.readme' file, which describes the various formats available to easily recover the data base on Macintosh computers. It should be pointed out that Colibri has been compiled as a 'clickable' Macintosh application, so that the 4D software is not needed to run it.

Users not connected to the Internet network can send five high-density (1.4 Mo) Macintosh diskettes to Secrétariat de Mr A. Danchin, Régulation de l'Expression Génétique Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cédex 15, France. However, the use of the network is strongly encouraged. We also encourage the report of any comment (and bugs) through electronic mail at bunny@radium.jussieu.fr.

APPENDIX

TABLE A1. Translation table for gene name equivalents

Alternate name	Name in Table 1	Alternate name	Name in Table 1	Alternate name	Name in Table 1	Alternate name	Name in Table 1	Alternate name	Name in Table 1	Alternate name	Name in Table 1	Alternate name	Name in Table 1
aceE1	*aceE*	*asp*	*ppc*	*cmlB*	*ompF*	*drm*	*deoB*	*fruC*	*fruR*	*herC*	*lysS*	*lexC*	*ssb*
aceE2	*aceF*	*aspB*	*gltD*	*cnt*	*cynS*	*drpA*	*proS*	*ftsB*	*nrdB*	*hid*	*himA*	*lky*	*tolA*
acrB	*gyrB*	*aspB*	*gltB*	*coa*	*ompF*	*dsf*	*ppfA*	*ftsI*	*pbpB*	*himB*	*gyrB*	*lon*	*capR*
ade	*purH*	*asuC*	*hisT*	*colE1*	*tolC*	*dtu*	*argD*	*fucC*	*fucA*	*himD*	*hip*	*lrp*	*livR*
ade	*purA*	*atp*	*cysC*	*con*	*ompA*	*eps*	*rpsE*	*gad*	*gap*	*hin*	*htpR*	*lspA*	*lsp*
ade	*purF*	*atpA*	*uncA*	*cou*	*gyrB*	*eryA*	*rplD*	*galA*	*galK*	*hisG*	*atpPT*	*lss*	*livR*
ade	*purN*	*atpB*	*uncB*	*cry*	*ompF*	*exbA*	*tonB*	*galB*	*galT*	*hlpA*	*skp*	*lstR*	*livR*
ade	*purC*	*atpC*	*uncC*	*csm*	*crp*	*excC*	*tolA*	*galD*	*galE*	*hom*	*asdX*	*malA*	*malT*
ade	*purB*	*atpD*	*uncD*	*ctr*	*ptsI*	*exrA*	*lexA*	*gcd*	*gldE*	*hpr*	*ptsH*	*malA*	*malP*
ade	*purE*	*atpE*	*uncE*	*ctr*	*ptsH*	*exrB*	*ssb*	*gen165*	*rsgA*	*hsm*	*hsdM*	*malA*	*malQ*
adth	*purD*	*atpF*	*uncF*	*cybA*	*sdhC*	*fabC*	*fabB*	*glc*	*crr*	*hsr*	*hsdR*	*malB*	*malK*
aidA	*alkA*	*atpG*	*uncG*	*cydA*	*cydI*	*fam*	*htpR*	*glgY*	*glgP*	*hss*	*hsdS*	*malB*	*malF*
aidD	*alkB*	*atpH*	*uncH*	*cydB*	*cydII*	*far*	*fusA*	*glmD*	*nagB*	*htrM*	*rfaD*	*malB*	*malG*
ala-act	*alaS*	*atpI*	*uncI*	*cysP*	*cysJ*	*fba*	*aldH*	*glnF*	*rpoN*	*htrP*	*luxH*	*malB*	*lamB*
alc	*fda*	*atpPT*	*hisG*	*cysQ*	*cysI*	*fba*	*fda*	*glnR*	*glnL*	*hu-2*	*hupA*	*malB*	*malE*
alt	*rpoD*	*azi*	*secA*	*cysZ*	*cysK*	*fbp*	*fdp*	*glnT*	*glnG*	*hycA*	*hevA*	*malM*	*molA*
ampA	*ampC*	*bfe*	*btuB*	*dam*	*damX*	*fdhA*	*selB*	*gltC*	*gltS*	*hycB*	*hevB*	*manX*	*ptsL*
amtA	*cysQ*	*bglB*	*bglC*	*dap*	*asdX*	*fdhA*	*selA*	*glu*	*ppc*	*hycC*	*hevC*	*manY*	*ptsP*
anth	*trpE*	*bglC*	*bglS*	*dapB*	*dapE*	*fdv*	*mutS*	*glut*	*gltA*	*hycD*	*hevD*	*manZ*	*ptsM*
apk	*lysC*	*bglF*	*bglC*	*dapX*	*nlpB*	*feuA*	*cir*	*glyD*	*glpD*	*hycE*	*hevE*	*mas*	*aceB*
arcA	*dye*	*bioB*	*bioH*	*dar*	*uvrA*	*feuB*	*fepA*	*gmk*	*spoR*	*hycF*	*hevF*	*mec*	*dcm*
arg	*carA*	*bioR*	*birA*	*dasF*	*rnh*	*fexA*	*dye*	*gpp*	*gppA*	*hycG*	*hevG*	*mel-4*	*melB*
arg	*carB*	*bisB*	*chlE*	*dda*	*uvrD*	*fii*	*tolQ*	*gpp*	*gpt*	*hycH*	*hevH*	*mel-7*	*melA*
arg1	*argA*	*blgA*	*bglB*	*dec*	*fadR*	*fipA*	*trxA*	*gpt*	*ptsG*	*icl*	*aceA*	*meoA*	*ompC*
arg1	*argD*	*blgG*	*bglS*	*deg*	*capR*	*flaBI*	*fliF*	*gptB*	*ptsM*	*iex*	*crr*	*metM*	*metL*
arg2	*argC*	*btuA*	*btuB*	*dhbB*	*birA*	*flaF*	*hag*	*gptB*	*ptsL*	*ihp*	*livR*	*mglP*	*mglB*
arg2	*argA*	*cap*	*crp*	*dhl*	*lpd*	*FlaJ*	*motB*	*groN*	*rpoB*	II-CAT	*cmlA*	*mglP*	*mglC*
arg4	*argE*	*cap*	*carA*	*dinA*	*polB*	*flaJ*	*motA*	*groNB*	*nusB*	*ind*	*tnaA*	*mglP*	*mglA*
arg5	*argF*	*cap*	*carB*	*dir*	*capR*	*flaN*	*fliE*	*groP*	*dnaJ*	IS*91*	*tnpA*	*micA*	*mutY*
arg6	*argG*	*car*	*ptsG*	*divA*	*ftsA*	*flbC*	*fliD*	*groP*	*dnaB*	*kac*	*kdpC*	*minB*	*minE*
arg7	*argH*	*cbr*	*fepA*	*dnaD*	*dnaC*	*flhC*	*flaI*	*groP*	*dnaK*	*kac*	*kdpD*	*minB*	*minD*
argA	*argE*	*cbt*	*fepA*	*dnaF*	*nrdA*	*flhD*	*flbB*	*grpA*	*dnaB*	*kac*	*kdpB*	*minB*	*minC*
argB	*argA*	*cer*	*btuB*	*dnaL*	*lig*	*fliC*	*hag*	*grpF*	*dnaK*	*kac*	*kdpA*	*mlpA*	*lppX*
argC	*argB*	*cheD*	*tsr*	*dnaP*	*dnaG*	*fliL*	*flaAI*	*gsa*	*popC*	*kdgA*	*ada*	*mopA*	*groEL*
argD	*argF*	*cheM*	*tar*	*dnaQ*	*mutD*	*fliM*	*flaAII*	*gshA*	*gshII*	*kga*	*ada*	*mopB*	*groES*
argE	*argG*	*cheX*	*cheR*	*dnaS*	*dut*	*flrD*	*fadI*	*gsr*	*crr*	*kgtP*	*witA*	*mor*	*oxyR*
argF	*argH*	*chlC*	*narH*	*dnaW*	*adk*	*folC*	*dedC*	*gurA*	*uidA*	*kps*	*neuS*	*mpt*	*ptsL*
argG	*argD*	*chlC*	*narG*	*dppA*	*fpp*	*fpk*	*fruK*	*gxu*	*gpt*	*leuK*	*hisT*	*mpt*	*ptsM*
argH	*argC*	*cim*	*tolA*	*dra*	*deoC*	*frr*	*rrfX*	*herA*	*rnh*	*lexB*	*recA*	*mra*	*murF*

Continued on following page

TABLE A1—*Continued*

Alternate name	Name in Table 1	Alternate name	Name in Table 1	Alternate name	Name in Table 1	Alternate name	Name in Table 1	Alternate name	Name in Table 1	Alternate name	Name in Table 1	Alternate name	Name in Table 1
mraY	*murX*	*nrdB*	*sodA*	*pea*	*secA*	*pyrA*	*carB*	*ron*	*rpoB*	*ssyF*	*rpsA*	*tpo*	*envZ*
mrcA	*ponA*	*ntrA*	*rpoN*	*pel*	*ptsP*	*rac*	*recE*	*rorA*	*recB*	*ssyG*	*infB*	*tpp*	*deoA*
mrcB	*ponB*	*ntrB*	*glnL*	*pepH*	*pepD*	*radB*	*recN*	*rpmE*	*katG*	*stl*	*rpoB*	*trkC*	*kefC*
mrdA	*pbpA*	*ntrC*	*glnG*	*perA*	*envZ*	*ramA*	*rpsD*	*rpoH*	*htpR*	*strA*	*rpsL*	*trpX*	*miaA*
mrdB	*rodA*	*nucR*	*deoR*	*pgsB*	*lpxB*	*Rarg*	*argI*	*rts*	*coaA*	*stv*	*rpoB*	*tryD*	*trpE*
msbB	*mlt*	*nupA*	*tsx*	*phoT*	*pstB*	*rbsP*	*rbsA*	*sbl*	*gutB*	*sud2*	*rpsD*	*tryE*	*trpD*
msp	*dye*	*nur*	*katF*	*phoT*	*phoU*	*rbsP*	*rbsC*	*sdrA*	*rnh*	*sueB*	*prfA*	*tryp*	*trpC*
msyA	*hns*	*oldA*	*fadA*	*phoT*	*pstA*	*rbsP*	*rbsB*	*secC*	*rpsO*	*suf*	*sulA*	*tryp*	*trpA*
mtcB	*tolC*	*oldB*	*fadB*	*phoT*	*phoB*	*rbsP*	*rbsD*	*seg*	*dye*	*sulB*	*ftsZ*	*tryp*	*trpB*
muc	*capR*	*oleR*	*fadR*	*pil*	*fimD*	*rbsT*	*rbsC*	*serT*	*divE*	*sun*	*rho*	*tsl*	*lexA*
mutR	*mutH*	*ompB*	*envZ*	*pil*	*fimC*	*rbsT*	*rbsA*	*sfiA*	*sulA*	*supK*	*prfB*	*tsp*	*prc*
mutU	*uvrD*	*ompB*	*ompR*	*pil*	*fimB*	*recE*	*racC*	*sfiB*	*ftsZ*	*supX*	*topA*	*tsu*	*rho*
mvrC	*EB*	*ompE*	*phoE*	*pilA*	*fimA*	*recL*	*uvrD*	*sfrA*	*dye*	*T6rec*	*tsx*	*ttr*	*fadL*
nalA	*gyrA*	*oriJ*	*racC*	*plsA*	*adk*	*refI*	*tolC*	*sfrB*	*hlyT*	*tabD*	*rpoB*	*tut*	*ompA*
nalC	*gyrB*	*pac*	*pga*	*pmi*	*manA*	*refII*	*cet*	*shl*	*fruR*	*tabD*	*rpoC*	*uar*	*prfA*
narC	*narG*	*panK*	*coaA*	*poaA*	*putA*	*relC*	*rplK*	*sin*	*rnh*	*tgl*	*ptsG*	*umg*	*ptsG*
narD	*chlD*	*papA*	*uncA*	*polC*	*dnaE*	*resA*	*polA*	*sof*	*dut*	*tgs*	*crr*	*umuA*	*lexA*
narE	*chlE*	*papB*	*uncD*	*popE*	*hemC*	*rf1X*	*prfA*	*sohA*	*prlF*	*thdB*	*fadR*	*umuB*	*recA*
narR	*narL*	*papC*	*uncG*	*ppfA*	*dsf*	*rf2X*	*prfB*	*spcA*	*rpsE*	*thrD*	*thrA*	*ura*	*carB*
narR	*narX*	*papD*	*uncB*	*prlA*	*secY*	*rfaH*	*hlyT*	*spf*	*polA*	*thyR*	*deoB*	*ura*	*carA*
ncf	*hemB*	*papE*	*uncH*	*prv*	*mutH*	*rfs*	*fliD*	*spr*	*lexA*	*thyR*	*deoC*	*uraP*	*upp*
neaA	*rpsQ*	*papG*	*uncC*	*pssA*	*pss*	*Rgal*	*galR*	*srlA*	*gutA*	*tif*	*recA*	*usg1*	*usg*
nhaA	*ant*	*papH*	*uncE*	*pstC*	*phoW*	*rglA*	*mcrA*	*srlB*	*gutB*	*tmrA*	*folA*	*uvm*	*umuC*
nicA	*nadA*	*par*	*ompC*	*pstS*	*phoS*	*rglB*	*mcrB*	*srlD*	*gutD*	*tnaR*	*tnaA*	*uvm*	*umuD*
nirA	*fnr*	*parB*	*dnaG*	*psuA*	*rho*	*rhaC*	*rhaR*	*srlM*	*gutM*	*tol-3*	*tolB*	*uvrF*	*recF*
nirD	*nirB*	*paxA*	*dcd*	*ptsN*	*nagE*	*rhaC*	*rhaS*	*srlQ*	*gutQ*	*tol-8*	*tolC*	*visB*	*ubiH*
nirR	*fnr*	*pcbA*	*gyrB*	*pup*	*deoD*	*rif*	*rpoB*	*srlR*	*gutR*	*tolF*	*ompF*	*xerA*	*argR*
nitA	*rho*	*pck*	*pckA*	*purE*	*purK*	*rimA*	*rpmH*	*ssaF*	*rpmH*	*tolG*	*ompA*	*xonA*	*sbcB*
nitB	*rpoB*	*pdeB*	*uvrD*	*purG*	*purM*	*rne*	*ams*	*ssp*	*sspB*	*tonA*	*fhuA*	*zab*	*recA*
nmpA	*phoS*	*pdeC*	*lig*	*purI*	*purL*	*rnsA*	*rna*	*ssp*	*sspG*	*TP*	*deoA*		
nmpB	*phoR*	*pdzA*	*rplT*	*pyrA*	*carA*	*rnsC*	*rho*	*ssyB*	*nusB*	*tpiA*	*tpi*		

ACKNOWLEDGMENTS

We thank F. Lisacek for her constructive comments on the manuscript and I. Moszer for corrections to Table 1.

We thank the society ORSAN for a fellowship to C.M. during part of the work. We also acknowledge grants from CNRS (GDR 1029) and the French Ministry of Research (91C0802).

REFERENCES

1. **Bachmann, B. J.** 1990. Linkage map of *Escherichia coli* K-12, edition 8. Microbiol. Rev. **54**:130–197.
2. **Barker, W. C., D. G. George, H. W. Mewes, and A. Tsugita.** 1992. GenBank. Nucleic Acids Res. **20**(Suppl.):2023–2026.
3. **Blake, R. D., and P. W. Hinds.** 1984. Analysis of the codon bias in *E. coli* sequences. J. Biomol. Struct. Dyn. **2**:593–606.
4. **Brewer, B.** 1990. Replication and the transcriptional organization of the *Escherichia coli* chromosome. p. 61–83. *In* K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.
5. **Burks, C., M. J. Cinkosky, M. Fischer, P. Gilna, J. E.-D. Hayden, G. M. Keen, M. Kelly, D. Kristofferson, and J. Lawrence.** 1992. GenBank. Nucleic Acids Res. **20**(Suppl.):2065–2069.
6. **Churchill, G. A., D. L. Daniels, and M. S. Waterman.** 1989. The distribution of restriction enzyme sites in *Escherichia coli*. Nucleic Acids Res. **18**:589–597.
7. **Daniels, D. L., G. Plunkett, V. Burland, and F. R. Blattner.** 1992. Analysis of the *Escherichia coli* genome: DNA Sequence of the region from 84.5 to 86.5 minutes. Science **257**:771–778.
8. **Diday, E.** 1971. Une nouvelle méthode en classification automatique et reconnaissance des formes: la méthodes des nuées dynamiques. Rev. Statist. Appl. **19**:19–33.
9. **Gouy, M., and C. Gautier.** 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. **10**:7055–7074.
10. **Higgins, D. G., R. Fuchs, P. J. Stoehr, and G. N. Cameron.** 1992. The EMBL Data Library. Nucleic Acids Res. **20**(Suppl.):2071–2074.
11. **Hill, C. W., and B. W. Harnish.** 1981. Inversion between ribosomal RNA genes of *Escherichia coli*. Proc. Natl. Acad. Sci. USA **78**:7069–7072.
12. **Hill, M. O.** 1974. Correspondence analysis: a neglected multivariate method. Appl. Statist. **23**:40–353.
13. **Karlin, S., and C. Macken.** 1991. Assessment of inhomogeneities in an E. coli physical map. Nucleic Acids Res. **19**:4241–4246.
14. **Knott, V., D. J. Blake, and G. G. Brownlee.** 1989. Completion of the detailed restriction map of the *E. coli* genome by the isolation of overlapping cosmid clones. Nucleic Acids Res. **17**:5901–5912.
15. **Knott, V., D. J. G. Rees, Z. Cheng, and G. G. Brownlee.** 1988. Randomly picked cosmid clones overlap the *pyrB* and *oriC* gap in the physical map of the *E. coli* chromosome. Nucleic Acids Res. **16**:2601–2612.
16. **Kohara, Y., K. Akiyama, and K. Isono.** 1987. The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. Cell **50**:495–508.
17. **Kröger, M., R. Wahl, G. Schachtel, and P. Rice.** 1992. Compilation of DNA sequences of *Escherichia coli* (update 1992). Nucleic Acids Res. **20**(supplement):2119–2144.
18. **Kunisawa, T., M. Nakamura, H. Watanabe, J. Otsuka, A. Tsugita, L. S. L. Yeh, D. G. George, and W. C. Barker.** 1990. *Escherichia coli* K12 genomic database. Protein Seq. Data Anal. **3**:157–162.
19. **Médigue, C., J. P. Bouché, A. Hénaut, and A. Danchin.** 1990. Mapping of sequenced genes (700 kbp) in the restriction map of the *Escherichia coli* chromosome. Mol. Microbiol. **4**:169–187.
20. **Médigue, C., A. Hénaut, and A. Danchin.** 1990. *Escherichia coli* molecular genetic map (1000 kbp): update I. Mol. Microbiol. **4**:1443–1454.

21. **Médigue, C., T. Rouxel, P. Vigier, A. Hénaut, and A. Danchin.** 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. J. Mol. Biol. **222**:851–856.
22. **Médigue, C., A. Viari, A. Hénaut, and A. Danchin.** 1991. *Escherichia coli* molecular genetic map (1500 kbp): update II. Mol. Microbiol. **5**:2629–2640.
23. **Miller, W., J. Ostell, and K. E. Rudd.** 1990. An algorithm for searching restriction map. Comput. Appl. BIOSci. **6**:247–252.
24. **Perrière G.** 1991. ColiGene, p. 61–66. *In* Traitement de l'information des séquences biologiques. Cahiers IMABIO no. 2. CNRS, Paris.
25. **Rechenmann, F., and P. Uvietta.** 1990. Shirka: an object-centered knowledge based management system, p. 9–23. *In* Intelligence artificielle en simulation numérique et symbolique. Aléas, Lyon, France.
26. **Rudd, K. E., W. Miller, J. Ostell, and D. A. Benson.** 1990. Alignment of *Escherichia coli* K12 DNA sequences to a genomic restriction map. Nucleic Acids Res. **18**:313–321.
27. **Rudd, K. E., W. Miller, C. Werner, J. Ostell, C. Tolstoshev, and S. G. Satterfield.** 1991. Mapping sequenced *E. coli* genes by computer: software, strategies and examples. Nucleic Acids Res. **19**:637–647.
28. **Shin, D. G., C. Lee, J. Zhang, K. E. Rudd, and C. M. Berg.** 1992. Redesigning, implementing and integrating *Escherichia coli* genome software tools with an object-oriented database system. Comput. Appl. BIOSci. **8**:227–238.
29. **Thierry-Mieg, J., and R. Durbin.** 1992. ACeDB—a *C. elegans* data base, p. 15–19. *In* Informatique et génomes. Traitement de l'information des séquences biologiques. Cahiers IMABIO no. 5. CNRS, Paris.
30. **Wilbur, W. J., and D. J. Lipman.** 1983. Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl. Acad. Sci. USA **80**:726–730.
31. **Yura, T., H. Mori, H. Nagai, T. Nagata, A. Ishihama, N. Fujita, K. Isono, K. Mizobuchi, and A. Nakata.** 1992. Systematic sequencing of the *Escherichia coli* genome: analysis of the 0–2.4 min region. Nucleic Acids Res. **20**:3305–3308.