Mapping sequenced *E. coli* genes by computer: software, strategies and examples

Kenneth E. Rudd*, Webb Miller¹, Craig Werner², James Ostell³, Carolyn Tolstoshev³ and Steven G. Satterfield³

Laboratory of Bacterial Toxins, Division of Bacterial Products, Center for Biologics Evaluation and Research, Food and Drug Administration, Bethesda, MD 20892, ¹Department of Computer Science, The Pennsylvania State University, University Park, PA 16802, ²Albert Einstein College of Medicine, New York City, NY 10461 and ³National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received August 15, 1990; Revised and Accepted December 7, 1990

**ABSTRACT**

Methods are presented for organizing and integrating DNA sequence data, restriction maps, and genetic maps for the same organism but from a variety of sources (databases, publications, personal communications). Proper software tools are essential for successful organization of such diverse data into an ordered, cohesive body of information, and a suite of novel software to support this endeavor is described. Though these tools automate much of the task, a variety of strategies is needed to cope with recalcitrant cases. We describe such strategies and illustrate their application with numerous examples. These strategies have allowed us to order, analyze, and display over one megabase of *E. coli* DNA sequence information. The integration task often exposes inconsistencies in the available data, perhaps caused by strain polymorphisms or human oversight, necessitating the application of sound biological judgment. The examples illustrate both the level of expertise required of the database curator and the knowledge gained as apparent inconsistencies are resolved. The software and mapping methods are applicable to the study of any genome for which a high resolution restriction map is available. They were developed to support a weakly coordinated sequencing effort involving many laboratories, but would also be useful for highly orchestrated sequencing projects.

**INTRODUCTION**

The complete DNA sequence of an organism’s genome is the entire set of instructions required for the assemblage and propagation of a cellular life form. The modeling of cellular structure and function based on a precise knowledge of the structural and regulatory information contained within genomic DNA is a formidable interdisciplinary task that will require the skills of biologists and computer scientists engaged in both practical and theoretical studies. A successful model of cellular growth and adaptation in response to external stimuli for any organism would greatly enhance our ability to understand and manipulate a variety of cells due to the many unifying concepts and structural similarities found among all life forms.

The bacterium *E. coli* offers the greatest hope for accomplishing such a task, since more is known about the structure and function of this cell than any other cell type. More than 1400 genes have been identified and mapped (1). In addition, the simplicity of its genome, 4.7×10⁹ bp of DNA organized into a single molecule (2), and the large number of laboratories currently engaged in sequencing *E. coli* genes suggest that *E. coli* may be the first organism whose genome is completely sequenced. Whereas achieving the monumental goal of sequencing the ~3×10⁹ bp of the human genome depends on the development of new technology, such is not the case for *E. coli*. In fact, over 25% of the *E. coli* genome has been sequenced to date as a collective, uncoordinated effort (3). Although three groups have each declared intentions of sequencing the entire *E. coli* genome themselves (4–6), no results of these efforts have yet been published. Moreover, it is unlikely that any single group would be able to provide the richness of information that usually accompanies publication of a single gene sequence. This information includes critical RNA and protein termini determinations as well as mutational and gene inactivation studies. Such information is vital to determining the accuracy and relevance of raw sequence data.

With these considerations in mind, we have developed a software system for collecting, aligning, and displaying *E. coli* genomic genetic map (1), restriction map (2), and DNA sequence information obtained in many different laboratories. Although

* To whom correspondence should be addressed
this mosaic has the disadvantage of involving DNA sequences from different strains of *E. coli* K12, it has the advantage that most sequences have been carefully determined, annotated, and confirmed in some way. This paper presents numerous examples of how new information has been obtained by manipulating published DNA sequence and restriction map data, including our digital version of the *E. coli* genomic restriction map (7). These examples are chosen to illustrate the function and utility of the new programs.

The methods we describe can be applied to the study of any genome for which a reasonable number of DNA sequences are already known and for which a high resolution genomic restriction map is available. Our software development is continuing and some of these approaches, although useful in the short term, have been superseded by more general or more portable approaches. The collection of ordered, non-overlapping *E. coli* DNA sequences obtained and analyzed using this software will be described in detail elsewhere (K.E.R. and W.M., manuscript in preparation).

**MAPPING SOFTWARE**

This section describes the new software, which includes refinements and extensions of our previously described software (7,8) as well as several completely new programs. DigiMap is a program for entering physical and genetic map data from hardcopy sources into the computer. PrintMap produces publication quality genomic restriction maps depicting the extents of other information aligned to the map, such as sequences, contigs, clones, genes, and transcripts, thereby permitting rapid visual correlation of overlapping data sets. MapSearch, our previously described restriction map alignment program, is now faster and more convenient to use. PrintAlign graphically depicts restriction map alignments produced by MapSearch. Finally, AlterMap uses a MapSearch alignment to splice a restriction map derived from a DNA sequence into the genomic restriction map. In addition, we custom-built a graphical interface that allows the user to select programs and options using a mouse.

In addition to the software that we developed, we utilized the GenInfo information retrieval software (9) and the DM5 DNA sequence manipulation and analysis software package (10). Overlaps of DNA sequence were detected using the FASTA homology searching routines (11).

**DigiMap**

DigiMap creates a digitized version of a genetic or physical map and stores it in a disk file. Maps to be digitized are in printed form and consist of a series of vertical or horizontal lines, with a marking for each gene or restriction enzyme recognition site, accompanied by a label for that site. A one kilobase or one minute scale bar can be digitized directly from the input figure or entered by typing a number. The digitization process may be completed in one session or by repeated invocations of the program, and existing computerized maps can be browsed and edited. DigiMap's graphical user interface presents a pictorial representation of the digitized map. User control is via a panel of action buttons using the mouse. The user is prompted to enter site addresses (via the digitizing tablet) and labels (via the keyboard).

We use DigiMap to enter and update genetic maps, including both the *E. coli* and *Salmonella typhimurium* maps. We have developed a method for producing high quality hardcopy versions of genetic maps using the Plasmid Description Language (see below) and data collected with DigiMap. Figure 1 depicts a portion of the 1983 *E. coli* genetic map (12) that has been digitized and displayed using this method. We prefer the 1983 map format over the 1990 map format despite the fact that we use the 1990 map data in Table 1. The 1990 map includes a number of expanded segments that are drawn with different scales. The uniform scale of the 1983 map makes it easier to assess the distance between genes as well as simplifying the digitizing process itself. Digital genetic maps can be easily updated and distributed using electronic media. DigiMap also allows rapid digitization of local physical maps which can be stored for later reference, transferred to others via electronic mail, updated or altered, converted to hardcopy using PrintMap, and used as MapSearch probes.

**PrintMap**

PrintMap prints a restriction map together with integrated cloning and sequencing information. For example, it can print an arbitrary segment of the *E. coli* genomic restriction map in a variety of styles and a wide range of resolutions (Figures 2 and 3). Its ASCII output is combined with the Plasmid Description Language (C.W., manuscript in preparation) and printed on a PostScript-compatible device. The program has been implemented under several versions of Unix, under IBM PC-DOS, and on the Apple Macintosh. On the IBM-PC, it can be combined with a PostScript emulator to produce output on a dot matrix printer or, with accessory programs, on a color terminal.

PrintMap is given the name of a map and the end points, measured in basepairs, of the region of the map to be printed. Foremost among numerous optional switches is a facility to specify files containing a number of lines of the form (starting_kb_address, ending_kb_address, label_text). Each line of the input file is depicted on the printed map as a 'span line' with the label_text underneath (see Figures 2 and 3). These span lines can be used to represent cloned segments of the genome, to display the extent of individual contiguous segments of DNA sequence or protein coding regions, or to demarcate special features such as a region of uncertainty, a gap, or an insertion (e.g., a prophage). When the map regions indicated by span lines overlap, a periodicity (number of span line 'steps') can be established for vertically staggering the span lines, and different sets of span lines can be given different periodicities.

Another common option determines whether the map is printed in an eight-line format like the one used by Kohara et al. (2), (e.g. Figure 3) or in an alternate format, where the map is represented as a single line, with tic marks of different heights representing the various enzyme sites, and with each tic mark labeled by a one letter code for the eight enzymes used (e.g., Figure 2). Other options affect margins, distances, spacing of numbers and of address calibration marks, number of lines the map contains on each page, placement and orientation of the map on the page, printing of the legend, and scaling of portions of the map to fill the available space. Finally, a complex set of option parameters can be stored in a file and later invoked with a single command, a procedure that saves time and preserves a record of PrintMap usage.

**MapSearch**

MapSearch determines a specified number of best alignments of a short 'probe' restriction map to regions of a longer genomic restriction map. Earlier papers (7,8) describe our method of defining and scoring alignments. Our technique, which we
Figure 1. Computer drawing of a portion of the E. coli genetic map. The positions and gene names of the genes in the first 50 minutes of the 1983 E. coli genetic map (12) were computerized using the DigiMap program. The map was formatted using the Plasmid Description Language (see text) and printed on a laser printer.

Our new version of MapSearch offers several improvements. First, it implements an algorithm that is appreciably faster for long probes. Its execution time is now proportional to M P log P rather than to M P squared, where there are M map sites and P probe sites (W.M., Barr, J., and K.E.R., manuscript submitted for publication). Moreover, it is now possible to request that specific restriction enzyme sites, e.g. EcoRV, be completely ignored. EcoRV sites are missing from certain regions of the genomic restriction map, and their presence in a sequence-derived restriction map used as a probe can cause misalignment when
TABLE 1. Sequence and restriction map alignments to the *E. coli* genomic map.

<table>
<thead>
<tr>
<th>Gene</th>
<th>bOri</th>
<th>bMin.</th>
<th>dRank</th>
<th>p value</th>
<th>fLocus</th>
<th>aAcc. #</th>
<th>hBp</th>
<th>hSites</th>
<th>jFirst</th>
<th>jLast</th>
<th>kDB</th>
<th>lType</th>
</tr>
</thead>
<tbody>
<tr>
<td>tdcC</td>
<td>-</td>
<td>68.300</td>
<td>1</td>
<td>&lt;0.001</td>
<td>ECTDCRAB</td>
<td>X14430</td>
<td>6295</td>
<td>9</td>
<td>3159000</td>
<td>3165600</td>
<td>E</td>
<td>repeats</td>
</tr>
<tr>
<td>tfeC</td>
<td>-</td>
<td>68.300</td>
<td>2</td>
<td>&lt;0.001</td>
<td>ECTDCRAB</td>
<td>X14430</td>
<td>6295</td>
<td>9</td>
<td>3172700</td>
<td>3179000</td>
<td>E</td>
<td>repeats</td>
</tr>
<tr>
<td>tfeC</td>
<td>-</td>
<td>68.300</td>
<td>3</td>
<td>&lt;0.001</td>
<td>ECTDCRAB</td>
<td>X14430</td>
<td>6295</td>
<td>9</td>
<td>3323900</td>
<td>3335800</td>
<td>E</td>
<td>repeats</td>
</tr>
<tr>
<td>tfeC</td>
<td>-</td>
<td>68.300</td>
<td>4</td>
<td>&lt;0.001</td>
<td>ECTDCRAB</td>
<td>X14430</td>
<td>6295</td>
<td>9</td>
<td>3145400</td>
<td>3152000</td>
<td>E</td>
<td>repeats</td>
</tr>
<tr>
<td>tfeA</td>
<td>-</td>
<td>68.300</td>
<td>10</td>
<td>0.996</td>
<td>tdcReco</td>
<td>ES8002</td>
<td>3000</td>
<td>3</td>
<td>3319700</td>
<td>3321800</td>
<td>ES</td>
<td>modified</td>
</tr>
<tr>
<td>IS5</td>
<td>ND</td>
<td>2</td>
<td>ND</td>
<td>INSSPEC</td>
<td>J01734</td>
<td>1300</td>
<td>2</td>
<td>3318100</td>
<td>3318800</td>
<td>G</td>
<td>repeat</td>
<td></td>
</tr>
<tr>
<td>fecA</td>
<td>ND</td>
<td>7.800</td>
<td>&lt;100</td>
<td>ND</td>
<td>ECOFEC</td>
<td>M20981</td>
<td>2645</td>
<td>7</td>
<td>ND</td>
<td>ND</td>
<td>G</td>
<td>repeat</td>
</tr>
<tr>
<td>fecB</td>
<td>+</td>
<td>7.800#</td>
<td>1</td>
<td>0.806</td>
<td>ECOFEC</td>
<td>M20981</td>
<td>2645</td>
<td>7</td>
<td>4589800</td>
<td>4591500</td>
<td>G</td>
<td>repeat</td>
</tr>
<tr>
<td>fecC</td>
<td>+</td>
<td>7.800#</td>
<td>1</td>
<td>1.000</td>
<td>M26397</td>
<td>M26397</td>
<td>4842</td>
<td>6</td>
<td>3254000</td>
<td>3287000</td>
<td>G</td>
<td>repeat</td>
</tr>
<tr>
<td>fecD</td>
<td>+</td>
<td>7.800#</td>
<td>1</td>
<td>&lt;0.001</td>
<td>M26397</td>
<td>M26397</td>
<td>4852</td>
<td>6</td>
<td>4585500</td>
<td>4589800</td>
<td>G</td>
<td>repeat</td>
</tr>
<tr>
<td>fecE</td>
<td>+</td>
<td>7.800#</td>
<td>6</td>
<td>1.000</td>
<td>fecEcoM</td>
<td>ES1033</td>
<td>7259</td>
<td>13</td>
<td>3254000</td>
<td>3335000</td>
<td>G</td>
<td>repeat</td>
</tr>
<tr>
<td>fecF</td>
<td>+</td>
<td>7.800#</td>
<td>1</td>
<td>&lt;0.001</td>
<td>fecEcoM</td>
<td>ES1033</td>
<td>7259</td>
<td>13</td>
<td>4585500</td>
<td>4592700</td>
<td>G</td>
<td>repeat</td>
</tr>
<tr>
<td>iap</td>
<td>+</td>
<td>59.200</td>
<td>9</td>
<td>1.000</td>
<td>ECOIAP</td>
<td>M18270</td>
<td>1664</td>
<td>3</td>
<td>2890700</td>
<td>2891600</td>
<td>G</td>
<td>normal</td>
</tr>
<tr>
<td>iap</td>
<td>+</td>
<td>59.600</td>
<td>1</td>
<td>0.141</td>
<td>iap-eco</td>
<td>ES8003</td>
<td>1665</td>
<td>4</td>
<td>2890300</td>
<td>2891600</td>
<td>ES</td>
<td>modified</td>
</tr>
<tr>
<td>bioH</td>
<td>-</td>
<td>75.000</td>
<td>15</td>
<td>1.000</td>
<td>ECOBIOH</td>
<td>X15587</td>
<td>2614</td>
<td>10</td>
<td>3613300</td>
<td>3615200</td>
<td>E</td>
<td>normal</td>
</tr>
<tr>
<td>bioH</td>
<td>-</td>
<td>75.000</td>
<td>15</td>
<td>0.011</td>
<td>bioHeco</td>
<td>ES3007</td>
<td>2537</td>
<td>4</td>
<td>3613300</td>
<td>3615200</td>
<td>ES</td>
<td>modified</td>
</tr>
<tr>
<td>purK</td>
<td>ND</td>
<td>12.200</td>
<td>1</td>
<td>0.287</td>
<td>ECOFEC</td>
<td>M19657</td>
<td>2449</td>
<td>2</td>
<td>5617000</td>
<td>5627000</td>
<td>G</td>
<td>twoise</td>
</tr>
<tr>
<td>pin</td>
<td>+</td>
<td>25.800</td>
<td>1</td>
<td>1.000</td>
<td>ECOHINP</td>
<td>K03521</td>
<td>2614</td>
<td>2</td>
<td>1220400</td>
<td>1221700</td>
<td>G</td>
<td>twoise</td>
</tr>
<tr>
<td>dnaC</td>
<td>-</td>
<td>98.900</td>
<td>2</td>
<td>0.575</td>
<td>ECOHINATC</td>
<td>J04030</td>
<td>2554</td>
<td>2</td>
<td>4678600</td>
<td>4679700</td>
<td>G</td>
<td>twoise</td>
</tr>
<tr>
<td>fic</td>
<td>-</td>
<td>74.200</td>
<td>58</td>
<td>1.000</td>
<td>ECOFIC1</td>
<td>M28363</td>
<td>2496</td>
<td>7</td>
<td>3560700</td>
<td>3562000</td>
<td>G</td>
<td>8enz</td>
</tr>
<tr>
<td>fic</td>
<td>-</td>
<td>74.200</td>
<td>58</td>
<td>1.000</td>
<td>ECOFIC1</td>
<td>M28363</td>
<td>2496</td>
<td>7</td>
<td>3560700</td>
<td>3562000</td>
<td>G</td>
<td>7enz</td>
</tr>
<tr>
<td>katG</td>
<td>+</td>
<td>89.150</td>
<td>4</td>
<td>0.966</td>
<td>ECOHINATC</td>
<td>M21516</td>
<td>2805</td>
<td>7</td>
<td>4212100</td>
<td>4213600</td>
<td>G</td>
<td>8enz</td>
</tr>
<tr>
<td>katG</td>
<td>+</td>
<td>89.150</td>
<td>4</td>
<td>0.966</td>
<td>ECOHINATC</td>
<td>M21516</td>
<td>2805</td>
<td>7</td>
<td>4212100</td>
<td>4213600</td>
<td>G</td>
<td>7enz</td>
</tr>
<tr>
<td>fruE</td>
<td>-</td>
<td>15.950#</td>
<td>1</td>
<td>0.010</td>
<td>ECOFUEHE</td>
<td>X17615</td>
<td>2900</td>
<td>4</td>
<td>1175300</td>
<td>1178500</td>
<td>E</td>
<td>normal</td>
</tr>
<tr>
<td>ptsG</td>
<td>+</td>
<td>24.400</td>
<td>1</td>
<td>0.008</td>
<td>ECOPTSG</td>
<td>J02618</td>
<td>1523</td>
<td>4</td>
<td>1174300</td>
<td>1175300</td>
<td>G</td>
<td>normal</td>
</tr>
<tr>
<td>gdhA</td>
<td>+</td>
<td>27.000#</td>
<td>43</td>
<td>ND</td>
<td>ECOGDHA</td>
<td>K02499</td>
<td>1937</td>
<td>2</td>
<td>1861700</td>
<td>1862400</td>
<td>G</td>
<td>7enz</td>
</tr>
<tr>
<td>gdhA</td>
<td>+</td>
<td>27.000#</td>
<td>1</td>
<td>0.168</td>
<td>gdhAecoP</td>
<td>EM6002</td>
<td>6800</td>
<td>8</td>
<td>1855700</td>
<td>1865800</td>
<td>EM</td>
<td>Pmap</td>
</tr>
<tr>
<td>crp</td>
<td>+</td>
<td>73.500</td>
<td>33</td>
<td>ND</td>
<td>ECOCRP</td>
<td>J01598</td>
<td>1127</td>
<td>4</td>
<td>3556300</td>
<td>3556800</td>
<td>G</td>
<td>normal</td>
</tr>
<tr>
<td>sodA</td>
<td>-</td>
<td>88.450</td>
<td>11</td>
<td>1.000</td>
<td>ECOSSOD</td>
<td>X09351</td>
<td>1053</td>
<td>3</td>
<td>4179800</td>
<td>4180100</td>
<td>G</td>
<td>normal</td>
</tr>
<tr>
<td>sodA</td>
<td>+</td>
<td>88.450</td>
<td>1</td>
<td>0.001</td>
<td>sodAecoP</td>
<td>EM6003</td>
<td>7250</td>
<td>6</td>
<td>4176200</td>
<td>4183400</td>
<td>EM</td>
<td>Pmap</td>
</tr>
</tbody>
</table>

(a) Gene contained in DNA sequence or restriction map. (b) Orientations of the aligned genes as determined by the MapSearch program. A plus (+) sign indicates that genes are transcribed in the direction of increasing genomic map coordinates (clockwise); a minus (-) sign indicates counterclockwise transcription. (c) The map position (in minutes). The positions are approximated from the 1990 *E. coli* genetic map (2). We realize that the genetic map positions were not originally determined to this level of accuracy but we imposed a resolution of 0.025 minutes to preserve map order information. Minutes marked with # do not agree with MapSearch alignments (see text). (d) The rank of the MapSearch alignment. (e) The p (probability) value calculated for the alignment using 100 map shuffles (7). (f) The name of the database entry. An M at the end of an EcoSeq database entry name denotes it is a meld of several sequences. Meld descriptions are available upon request. A P at the end of an EcoMap database entry name denotes it is a probe derived from a published restriction map. (g) Database accession numbers of files which contain detailed information used in this analysis. (h) The number of basepairs represented by MapSearch probes. (i) The number of restriction sites in each MapSearch probe. (j) The genomic addresses (in basepairs) of the first and last restriction sites aligned. (k) The database containing the sequence or restriction map information: G, GenBank; E, EMBL; ES, EcoSeq; EM, EcoMap. (l) The type of alignment used (see text): twoise, probes have only two sites; 7enz, EcoRv information is ignored; Pmap, a published physical map is used as a probe. ND, no data.

alignments as the user requests in command line arguments. These alphanumeric alignments indicate unaligned sites in either the probe or the genomic map. Also, the genomic map coordinates corresponding to the ends of the probe for each alignment are printed. These coordinates can be used by
AlterMap to splice an exact copy of a sequence-derived probe into the E. coli map, or used to produce a PrintMap span file that depicts sequenced regions.

MapSearch is a key component of our computerized E. coli genome research. The results of aligning individual restriction maps to the E. coli genomic map by eye have been reported by many researchers. We have found that this process can easily produce biased results and offers no statistical basis by which reliability can be assessed. Now, anyone with an IBM-compatible, Apple, Unix, or VAX/VMS computer can utilize MapSearch to align restriction maps to the genomic restriction map in a comprehensive and systematic fashion.

PrintAlign
PrintAlign displays a MapSearch alignment, giving either an alphanumeric representation or a line drawing. A printed alignment can be displayed on an ASCII printer or terminal. Line drawings of alignments are displayed on a Postscript laser printer or on a UNIX workstation running the X11 windowing system with a Tektronix terminal emulator. An example of PrintAlign output is given in Figure 5. An alternative labeling scheme places the name and address at the physical location of the site. For probes with closely spaced sites, the names may overlap, making them difficult to read.

PrintAlign has enabled us to identify putative insertions and deletions when comparing two restriction maps. These graphic representations of map alignments are interpreted much more easily and quickly than the alphanumeric representations that MapSearch produces.

AlterMap
AlterMap uses a MapSearch alignment to replace an appropriate section of a genomic map by a piece that exactly matches the restriction map used as a probe. For example, one might save the MapSearch alignments in a file, determine that the second highest scoring alignment is the correct one, and instruct AlterMap to alter the region of the map delimited by the chosen alignment to exactly match the probe. AlterMap automatically updates the map’s recorded length (which is always on the first line of a map file) and adjusts addresses of map sites that follow the changed portion by adding or subtracting an appropriate multiple of 100 bp. Optionally, files containing PrintMap span-line information are simultaneously updated. Since our digitization of the E. coli genomic map rounded restriction site addresses to the nearest multiple of 100 bp, one can tell at a glance which regions of the modified map are not DNA sequence-derived.

We have used AlterMap to replace 27.4% of the genomic restriction map with sequence-derived (and presumably more accurate) restriction map information (K.E.R., manuscript submitted for publication). This was done using 252 DNA sequence-derived probes (representing a total of 1283.7 kb of sequence information). A batch procedure alternately invokes MapSearch and then AlterMap for each probe. Thus each probe is aligned to a map that has been altered by all of the preceding probes. The total genome length was increased by 8.9 kb (0.7%) using this procedure, indicating that the partial restriction enzyme digestion method used to produce the genomic restriction map (2) was unexpectedly accurate in estimating the size of the E. coli genome.

**MAPPING STRATEGIES AND EXAMPLES**

We have continued to use MapSearch to align individual GenBank and EMBL sequence entries to the genomic restriction map of E. coli essentially as described earlier (7). This section begins with two particularly interesting examples that arose during routine applications of MapSearch; both examples concern the discovery of putative duplicate DNA sequences. We then describe several strategies that have been developed to handle sequences...
containing too little information to be aligned directly by MapSearch.

MapSearch Examples

Occasionally, routine use of MapSearch produces a result that suggests the existence of duplicate copies of a gene. Table 1 includes the top four MapSearch alignments for the GenBank sequence entry ECOTDCRAB(15) encoding the tdc (threonine dehydratase) operon, which was not expected to be present in multiple copies. The region containing these highly significant alignments has previously been shown to contain duplications of a 14 kb genomic region, with multiple copies of the ISS insertion sequence flanking the duplicated regions (16,17). These extra copies are absent from most E. coli strains, but present in the version of E. coli W3110 used to derive the genomic restriction map (17,18). The alignment at 3329.3 kb is consistent with the genetic map location and probably spawned the three copies located between 3145.4 kb and 3179.0 kb. Thus the three tandem copies have been removed from the current edition of our digital genomic restriction map since they are not present in other strains of E. coli K12, reducing the genome length estimate by 40.3 kb. Using either the first 3000 basepairs of ECOTDCRAB or a restriction map probe (see below) taken from the unsequenced part of the 14 kb duplicated region (called tdcReco in Table 1), a fifth (partial) copy of this duplicated region is located adjacent to the presumed original copy of the tdc operon. This fifth copy of the duplicated region was recently identified by eye and shown to be bracketed by ISS elements, as depicted in Figure 2 (17). A plausible mechanism by which the multiple copies at 3150 kb could have been spawned from the 3330 kb region duplication has been proposed, but the genesis of the duplication at 3330 kb, a region probably devoid of ISS elements in most other strains, remains unexplained (17). To change this region of the map in order to more accurately represent the wildtype E. coli chromosome, we have also removed the partial copy (11 kb) from our current version of the genomic restriction map.

MapSearch alignments suggest that the partial copy of the duplication (3319.7 to 3327.7, see Figure 2) may have resulted from an insertion of ISS into the tdc operon (in or near the tdcA gene). This event could have been accompanied by the loss of the distal tdc operon genes, recombination with a duplicate intact copy of the tdc operon and upstream region, possibly on a sister chromosome, and another transposition of ISS. This complicated series of events might have involved recombination between multiple copies of this region present during growth in rich media, since this region is not far from the origin of DNA replication. A frenzy of recombination may have been caused by high concentrations of ISS transposase due to the large number of insertion elements present in W3110. Whatever sequence of events did occur, it appears that the tdc operon was also inverted since this region is not far from the origin of DNA replication. Events did occur, it appears that the tdc operon was also inverted, but the genesis of the duplication at 3330 kb, a region probably devoid of ISS elements in most other strains, remains unexplained (17). To change this region of the map in order to more accurately represent the wildtype E. coli chromosome, we have also removed the partial copy (11 kb) from our current version of the genomic restriction map.

MapSearch examples suggest that the partial copy of the duplication (3319.7 to 3327.7, see Figure 2) may have resulted from an insertion of ISS into the tdc operon (in or near the tdcA gene). This event could have been accompanied by the loss of the distal tdc operon genes, recombination with a duplicate intact copy of the tdc operon and upstream region, possibly on a sister chromosome, and another transposition of ISS. This complicated series of events might have involved recombination between multiple copies of this region present during growth in rich media, since this region is not far from the origin of DNA replication. A frenzy of recombination may have been caused by high concentrations of ISS transposase due to the large number of insertion elements present in W3110. Whatever sequence of events did occur, it appears that the tdc operon was also inverted since this region is not far from the origin of DNA replication. Events did occur, it appears that the tdc operon was also inverted, but the genesis of the duplication at 3330 kb, a region probably devoid of ISS elements in most other strains, remains unexplained (17). To change this region of the map in order to more accurately represent the wildtype E. coli chromosome, we have also removed the partial copy (11 kb) from our current version of the genomic restriction map.

Figure 3. Restriction maps of the E. coli chromosome. A. The E. coli genomic restriction map from 275.0 to 355.0 kilobases is depicted. The fec operon (fecEcoM) location consistent with the genetic mapping data (1,21,22, see text) is in a region containing several insertion sequences (16,22,23) and a duplicate gene, argF (ECOARGF, 62). B. The E. coli genomic restriction map from 4550.0-4630.0 kilobases is depicted. This position of a possible second copy of the tdc operon (fecEcoM) is the highest-ranked MapSearch alignment (see Figure 5). The DNA sequences of the nearby genes argF (duplicate gene, ECOARGF, 62), tdcA) (the gene adjacent to tdcR (that encodes the RecA protein, perhaps the RecA protein, was involved, in the process (19, see Table 1). Finally, it is conceivable that a DNA binding protein, perhaps the RecA protein, was involved, since we have found that ISS and tdcR (the gene adjacent to tdcA also contain a DNA sequence of 18/21 identical bp (data not shown).
fecA fecBCDE). (Melding is discussed more thoroughly below.) The elements may have been involved in a putative translocation of elements (17, 23, 24) and the duplicate genes (argl/argF) and the sequences differ by only three base pairs (21).

The melded sequence aligns rather weakly to the 7.8 minute region as in reference 8. Orientation and p value are as in Table for homologous recombination. Moreover, as depicted in Figure 4 the genomic restriction map of this region (1). Some of the also caused a deletion in the clones that were used to produce genes (21), which may have sequence in the vicinity of the fe

Our second example concerns a discovery we made after creating a 'melded' fec operon of E. coli from the overlapping GenBank sequence entries ECOFEC (20, fecA) and M26397 (21, fecBCDE). (Melding is discussed more thoroughly below.) The genes have been genetically mapped to 7.8 minutes, but these mapping experiments are fraught with many inconsistencies and inexplicable results, and appear to be very unreliable (20, 22). The melded sequence aligns rather weakly to the 7.8 minute region (rank = 6, p = 1.0). Strikingly, the highest ranking alignment is at another locus, near 97.3 minutes (see Table 1 and Figure 3), as are the highest ranking alignments for probes derived from the ECOFEC and M26397 DNA sequences without melding (Table 1). In fact, for M26397 and the meld sequence fecEecoM, these are the only significant alignments, and they form nearly perfect matches (p < 0.001). Because of the high significance of the 97.3 minute alignment, it is possible that this is the region that has been cloned (20, 21). The ECOFEC (fecA) DNA is cloned from E. coli B because the cloned E. coli K12 fecA gene produced a truncated gene product. However the fecB region has been cloned from both E. coli K12 and E. coli B, and the sequences differ by only three base pairs (21).

We believe there may be two fec operons, one copy at 97.3 minutes and another copy at 7.8 minutes. The truncated fecA gene may well be the result of a cloning artifact due to an unclonable sequence in the vicinity of the fec genes (21), which may have also caused a deletion in the clones that were used to produce the genomic restriction map of this region (1). Some of the inexplicable mapping data could result from two target regions for homologous recombination. Moreover, as depicted in Figure 3, the DNA surrounding these two regions contains many IS elements (17, 23, 24) and the duplicate argl/argF genes (1) (argF is approx. 30 kb clockwise of the 7.8 minute locus and argl is approx. 30 kb clockwise of the 97.3 minute locus). The IS elements may have been involved in a putative translocation of the fec operon. Figure 3 depicts the arrangement of genes in the two fec regions. We note that apparently due to a typographical error, the fecA map position is given as 93 minutes in Table 1 of the E. coli genetic map (1).

Three Simple Variations

We have modified some GenBank and EMBL sequences so that they accurately represent the E. coli chromosomal DNA sequence. In some cases, the database entry does not contain all six bases of a restriction site known to be at one end. For example, we have added one G residue to the beginning of the GenBank entry ECOIAP (25) to complete the EcoRI site at the beginning of this sequence, thus forming the entry lap-eco. As seen in Table 1, this additional EcoRI site improves the rank of the correct MapSearch alignment from ninth to first. We have also found several entries that contain vector DNA sequence at the ends. For example, the EMBL entry ECBIOH (26) contains phage M13 polylinkers at both ends. We removed these vector sequences to create the entry bioHeco. This improves the correct MapSearch alignment from a rank of 15 to the best ranked alignment (see Table 1).

In our previous study (7, 8) we excluded the use of probes containing fewer than three restriction enzyme sites. We have since discovered that some two-site probes can indeed be aligned with MapSearch (although in some cases a statistical analysis is precluded). This alignment is possible because MapSearch uses the information between the ends of the probe and the terminal restriction sites. Some two-site probes are long enough that regions of the map with a correspondingly sparse distribution of sites are relatively rare, thus allowing the probe to be aligned. Table 1 contains three examples of such alignments: ECOPUREK (27) (best alignment), ECOPINP (28) (best alignment), and ECOCDNATC (29) (second best alignment). We now routinely use two-site probes with MapSearch.
The original *E. coli* genomic restriction map (2) contains 40 regions in which no EcoRV site information is given. Thus for 13% of the chromosome the genomic restriction map is comprised of sites for only seven restriction enzymes. However, our sequence-derived probes contain sites for all eight restriction enzymes. If the correct position for a DNA sequence lies in one of these regions lacking EcoRV information, then MapSearch may fail to correctly align the sequence. Therefore we have included the new version of MapSearch the ability to disregard any enzyme not included in an enzyme list. Probes are now routinely run through MapSearch with an enzyme list that does not include EcoRV, which has improved the rank and p value of many alignments to the regions without EcoRV information.

Two examples of EcoRV-less searches are given in Table 1. We were able align the GenBank DNA sequence ECOIFIC1 (30) to the genomic restriction map only by ignoring the EcoRV sites in the probe. The ECOIFIC1 probe contains seven restriction sites, three of which are EcoRV sites. An alignment consistent with the genetic map location of 74.2 minutes (1) is obtained as the top ranked alignment (p = .158) if the EcoRV sites are disregarded, whereas no suitable alignment is obtained if we search for recognition sites of all eight restriction enzymes (see Table 1). The good alignment falls within one of the genomic map regions without EcoRV sites. In another case, a probe derived from the GenBank DNA sequence ECOKATGA (31) contains seven sites, two of which are EcoRV sites. As shown in Table 1, if MapSearch looks for all eight enzymes, the correct alignment has a rank four and p value 0.966. If the EcoRV sites are ignored, the top ranked alignment is correct, with a p value of 0.191. The sequence ECOCOPDB (32) provides an example of how genes can be misaligned due to the lack of EcoRV information in the genomic map. We previously placed ECOCOPDB at 4526.9 kb by mistake, although we did report that the alignment lacked statistical significance (7). Using a published restriction map (32), we now correctly match the first aligned restriction site in ECOCOPDB with a genomic PvuII site at 4516.1 kb (data not shown). The ECOCOPDB sequence contains a total of only three sites, two EcoRV sites and one PvuII site. The sequence was misplaced earlier because the correct location (4516.1 kb) lacks EcoRV site information. In this case, ignoring the EcoRV sites in the probe doesn't help because MapSearch cannot utilize a one-site probe.

**MapSearch with Sequence Melds**

Although many small DNA sequences contain too few restriction sites to allow proper alignment with MapSearch, some of these can be combined with overlapping sequences to make a DNA sequence from which a longer probe can be created, as illustrated above with the fec operon. In some cases, a poorly aligned or unaligned sequence may meld to adjacent sequences that have been aligned successfully by MapSearch, usually improving the alignment of the DNA sequences. At other times, two genes that are not aligned by themselves can be aligned after being melded. Moreover, the process of finding sequence overlaps as an aid to aligning probes is sometimes reversed; MapSearch alignments that are adjacent on the genomic map occasionally reveal sequence overlaps that are very short. As an extreme case, a flush meld (no sequence overlap) may be suggested by MapSearch alignments. Examples are given below.

For many uses of our software, it is critical to meld as many *E. coli* DNA sequences as possible. For example, this is important for efficient use of AlterMap, and is essential for creating a database of non-overlapping *E. coli* DNA sequences to be used in sequence tallies (i.e., determining the percent of the genome that is sequenced). Finally, meld alignments are required in order to accurately annotate the integrated genomic map with sequence features such as translation and transcripts signals, a process we are currently completing. It should be noted that the problems associated with the lack of EcoRV information for 13% of the genome discussed above also occur with melded sequences.

The tyrT gene sequence, ECOTGY1 (33,34), provides an example of melding a poorly aligned sequence to adjacent sequences that are already aligned successfully by MapSearch. Our previous MapSearch alignment of ECOTGY1 was poor (7) and resulted in a clockwise orientation, as has been indicated in the genetic map (1). However, when melded to the adjacent nar gene sequences ECONARGLX (35), ECNARK (36), ECNARGHJ (37), and ECNARI (38), it becomes clear that the tyrT gene is actually transcribed in the counterclockwise direction. This map correction and numerous other examples of corrections to the *E. coli* genetic map (1) will be published in detail elsewhere (K.E.R. and W.M., manuscript in preparation).

The DNA sequences ECOPTS (39) and ECCHUWE (40) individually yield significant highest-ranked alignments with MapSearch (Table 1). While the ECOPTS alignment is consistent with the ptaG genetic map position (24.700 minutes), the ECOHUE sequence was significantly aligned to only one position, far from the reported flusE genetic map position of 15.950 minutes (1,41). The mapping of flusE to 16 minutes was reported to have been difficult and the expected linkage to some genes in the 16 minute region (iolA, kdp) was not detected (41). The two alignments overlap on the genomic map, suggesting DNA sequence overlap. This is indeed the case as the two sequences overlap by 297 nucleotides. Thus the flusE genetic map position is corrected to 24.800 minutes. The significance of the alignments is improved by the melding process (Table 1).

A routine search for sequence overlap between genes in the 70 minute region of the chromosome revealed that the ssp gene (69.950 minutes) sequence (ECOSSPG, 42) and the rpsl gene (70.250 minutes) sequence (ECORPSI, 43) are adjacent to one another. The first 223 base pairs of ECORPSI are identical to bases 892 to 1114 of ECOSSPG. However the last 70 bases of ECORPSI are not found in the ECOSPG sequence. These bases were found to be identical to the first 70 bases of the IS5 insertion sequence (17). The IS5 sequence located 63 basepairs after the end of the rpsl gene is not present in most strains of *E. coli*, and its presence here has been previously noted, although the overlap with ECOSPG was not previously detected (17). This DNA sequence meld, ssp-eacoM, refines the gene order in this region of the chromosome (1), removes anomalous DNA from the ECORPSI sequence, and allows MapSearch to align these sequences to the genomic restriction map. The two sequences contain only one restriction site each and could not be aligned by themselves. However the meld was aligned as the fifth best alignment using the two-site meld probe (Table 1).

The two DNA sequences ECOHEMA (44) and ECOHUE (45), containing the pf4A gene, have 570 bp of sequence overlap and were melded to form the sequence we call hemAecoM. This meld is significantly aligned, as is ECOPRS (46), to a position near its map position of 26.700 minutes (Table 1). The last restriction site aligned for ECOPRS is a BamHI site at 1275.2 kb and the first restriction site aligned for hemAecoM is the same BamHI site. Therefore, despite the fact that there is no sequence overlap other than the BamHI site itself, we propose that these sequences be combined into one large meld. A similar convergence of the MapSearch alignments of the sequences

---

**Table 1**

<table>
<thead>
<tr>
<th>DNA Sequence</th>
<th>MapSearch Alignment</th>
<th>Genetic Map Position</th>
<th>Corrected Genetic Map Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECOIFIC1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECOKATGA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECOCOPDB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECOPTS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECOHUE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECOSSPG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECORPSI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECOHEMA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECOHUE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECOHUE</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ECOVALS (47) and ECXERB (48, containing the pepA gene) on a common terminal HindIII site at 4559.5 kb was also observed (Table 1, Figure 3B). However, in this case the formation of a meld is confirmed by the alternative GenBank w15 sequence entry ECOSYNTGV (49), which contains 12 additional basepairs identical to the 12 ECXERB bases proximal to the common HindIII site (data not shown).

**MapSearch with Physical Maps as Probes**

Despite the use of two-site probes, EcoRV-less searches, and sequence melds, we were still unable to align many small DNA sequences to the genomic restriction map. Most of these remaining sequences have been aligned using probes derived from published restriction maps of unsequenced DNA flanking the sequenced region. A library of such probes has been assembled using DigiMap. MapSearch is usually applied to these probes with no enzyme list specified; MapSearch then recognizes only restriction sites that are present in the probe and ignores all others. Typically, only a few of the eight enzymes used to construct the genomic restriction map are present in these local restriction maps.

Three examples of this strategy are given in Table 1. Two entries are given for each case: one documents the poor performance of the sequence-derived restriction map probe; the second shows the improved results obtained when physical maps are used as probes. The ECOGDKHA (50) sequence contains five sites, three of which are EcoRV sites. Using MapSearch with or without the EcoRV sites, no reliable alignment could be found corresponding to its published genetic map location at 27.0 minutes (1). Therefore we constructed a physical map probe from the published restriction map (51). Table 1 shows that a very significant top alignment was located near 39 minutes. This is a region without EcoRV information, and using the sequence-derived probe without EcoRV, the location appeared as only the 43rd best alignment. Recently, the true map location of the gdhA gene has been shown to be 38.6 minutes (52), confirming our MapSearch alignment. Interestingly, the Salmonella typhimurium gdhA gene is located at 27.3 minutes (53). The E. coli and S. typhimurium genomes have been observed to be inverted with respect to one another between 25 minutes and 37 minutes (54) and perhaps this inversion extends to 38.6 minutes. In another case, the sequence ECOCR (55) contains two sites, and its probable location was found by MapSearch as the alignment of rank 33, which is too low to be considered as reliable. Using the published restriction map for the crp gene (55) as a MapSearch probe, the correct location was identified as the top-ranked alignment. In the third example, ECOSOD (56), the physical map (57) MapSearch alignment was used to confirm a sequence-derived MapSearch alignment with a low rank (=11th) and p value (1.0) (Table 1). Many more genes, both sequenced and unsequenced, have been aligned to the genomic restriction map easily and reliably using the physical maps and MapSearch (K.E.R. and W.M., manuscript in preparation). We routinely use this method to confirm any sequence-derived MapSearch alignments with low rank and insignificant p values.

**DISCUSSION**

We describe extended applications of our improved restriction map alignment program, MapSearch. Our examples illustrate how new genomic map information can be obtained using published information. In addition, we report the development of several new programs that are used in conjunction with MapSearch. Taken together, these programs are used to coordinate a genomic DNA sequencing project as a collective effort with contributions from many different laboratories. The methods we describe can be applied to the study of any genome for which a reasonable number of DNA sequences are known and for which a high resolution genomic restriction map is available.

Several distinct methods for aligning DNA sequence to the genomic restriction map have been described. With them, we have aligned nearly all of the >1.4 megabases of sequenced E. coli to the genomic restriction map. First, as previously published (7), DNA sequences are taken from primary sources, converted to restriction maps, and aligned directly with MapSearch. However, many DNA sequences are not aligned by this method, either because they contain an insufficient number of restriction sites to be aligned (even given a probable location from genetic map data), or because the genomic restriction map data is too inaccurate. Therefore, two additional approaches have been taken, both utilizing MapSearch and the other programs described herein. In one alternative, contiguous, overlapping DNA sequences are combined to create a longer sequence with more information content than the parent sequences. These melds can often be aligned using MapSearch. Occasionally this also fails to produce a good alignment, although part of the sequence aligns well by itself. A careful examination of several of these cases showed that the genomic restriction map disagrees with the sequence-derived data. Several sources may account for these discrepancies. Strain differences may be extensive due to reassortment of insertion sequences or other gross rearrangements. Cloned DNA may be deleterious to cell growth or vector replication, thus providing a selection for deletions of cloned DNA. One possible example of this is seen in the fecA region of the E. coli chromosome (see above). Also, the lack of EcoRV site information for a significant portion of the genomic map can cause misalignment (see above). We have aligned the sequences in question by aligning a neighboring DNA sequence to which it was melded. Our third method of sequence alignment uses neighboring restriction sites in unsequenced DNA to align DNA sequences that do not contain restriction sites or to bolster low-ranked or statistically insignificant alignments. The program DigiMap is used to enter this neighboring restriction site data, as illustrated by examples given above.

An alternative approach to restriction map alignments has recently been published (14). This method involves a search for regions of the genomic restriction map that contain restriction fragments whose sizes are consistent, within fixed, arbitrary limits, with those predicted from DNA sequences. No methods for ranking the matches or determining their statistical significance were described. Although insufficient data were presented to allow comparative performance evaluation, we feel that MapSearch provides the more rational and quantitative approach to restriction map alignments.

We stress that computerized alignments are not meant to substitute for direct experimentation to map genes. Rather, they are used in conjunction with genetic and physical mapping techniques. As described earlier (8), we utilize genetic map information to help decide if a particular alignment is the correct one. The ordered set of bacteriophage lambda clones that were used in determining the genomic restriction map are readily available to any researcher who desires them. In this way, alignments predicted by MapSearch can be easily confirmed, either by genetic complementation or, preferably, DNA hybridization experiments. In some cases, such as the tRNA genes...
Anonymous ftp protocol. Users connected to the Internet can obtain source codes for various programs using the MapSearch methods because they are too small and no restriction map is available. In these cases, the clones used to make the genomic restriction map contain the sequences of interest, as well as the sizes of the restriction fragments that contain the sequences, have been identified. This was enough information to allow the sequences to be positioned on the genomic map even though some of the sequences contained no restriction sites.

The alignment procedures we utilize serve to finely map sequenced genes relative to specific genomic restriction sites, to orient the genes relative to the direction of chromosome replication or the genetic map, and to organize the sequences in a logical fashion. This organization, particularly in its graphic PrintMap representation (see Figs 2 and 3), permits approximation of the distances between sequenced sections, often quite small. It also points out areas of inconsistent sequence or restriction map data. In order to expand the utility of MapSearch we have developed a program that looks for inverted and direct repeats in the restriction map patterns (W.M. and K.E.R., unpublished results). The quantitative aspects of MapSearch alignments can then be used to provide a measure of evolutionary relationships between homologous DNA segments. This method can also be used to compare the genomes of closely related species.

We have made available to the research community the digital E. coli restriction map, a set of non-overlapping DNA sequences in excess of 1.3 megabases, (K.E.R. and W.M., manuscript in preparation), and the genomic map positions and orientations of aligned DNA sequences. In addition, we provide computer files of clone endpoints, the no-EcoRV regions, digital versions of published local restriction maps, the orientations of cloned segments with respect to the arms of the lambda cloning vectors, and versions of the genomic restriction map containing sequence-derived restriction map information as produced by the AlterMap program. This set of data files will be updated frequently as new information is obtained. We are currently organizing most of this information into a relational database (G. Bouffard, J.O., and K.E.R.). We invite researchers to provide us with new E. coli DNA sequence and restriction map information; we will attempt to align this DNA to the genomic restriction map, ensuring confidentiality if requested. Researchers are also encouraged to send DNA sequences promptly to either the GenBank or EMBL repositories.

Upon receipt of the DNA sequence and restriction map information we will attempt to align the DNA to the genomic map, ensuring confidentiality, to obtain a measure of evolutionary relationships between homologous DNA segments and to orient the genes relative to the direction of chromosome replication. This method can also be used to compare the genomes of closely related species.

We have made available to the research community the digital E. coli restriction map, a set of non-overlapping DNA sequences in excess of 1.3 megabases, (K.E.R. and W.M., manuscript in preparation), and the genomic map positions and orientations of aligned DNA sequences. In addition, we provide computer files of clone endpoints, the no-EcoRV regions, digital versions of published local restriction maps, the orientations of cloned segments with respect to the arms of the lambda cloning vectors, and versions of the genomic restriction map containing sequence-derived restriction map information as produced by the AlterMap program. This set of data files will be updated frequently as new information is obtained. We are currently organizing most of this information into a relational database (G. Bouffard, J.O., and K.E.R.). We invite researchers to provide us with new E. coli DNA sequence and restriction map information; we will attempt to align this DNA to the genomic restriction map, ensuring confidentiality if requested. Researchers are also encouraged to send DNA sequences promptly to either the GenBank or EMBL databases, even if publication of the sequence is not anticipated. We will provide MapSearch, AlterMap, and PrintMap C language source code compatible with DOS, MacOS, Unix, and VMS operating systems to any interested parties. Data and programs are available online through anonymous ftp (see below). We hope that the E. coli research community will use the EcoSeq (E. coli DNA sequence database) and EcoMap (restriction map alignments database) resources as a clearinghouse for E. coli genome information and we will regularly assess and report on the progress of the collective E. coli genomic sequencing efforts. We strongly encourage all researchers to obtain access to electronic network facilities if possible, but we will also provide data and programs via IBM or Macintosh diskettes as requested.

A number of data files describing the E. coli genome and the source codes for various programs can be obtained using the anonymous ftp protocol. Users connected to the Internet can type 'ftp ncbi.nlm.nih.gov', enter 'anonymous' as user ID and use any word as password to access the repository/EcoSeq and repository/EcoMap subdirectories. Desired files are transferred to the user's computer with the command 'get filename'. Online information describing the content and format of the various files is available. Please address all electronic communications to 'rudd@ncbi.nlm.nih.gov'.

ACKNOWLEDGEMENTS

We would like to thank D. Benson and D. Lipman for their advice and encouragement, T. Kazic for helpful criticisms of the manuscript and W. Epstein for communicating results prior to publication. We are grateful for numerous discussions with B. Baum about the accuracy of E. coli DNA sequence and genetic mapping data. This publication was supported in part (W.M.) by grant RO1 LM5110 from the National Library of Medicine.

REFERENCES