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

Kenneth E.Rudd*, Webb Miller[1], Craig Werner[2], James Ostell[3], Carolyn Tolstoshev[3] and Steven G.Satterfield[3]

Laboratory of Bacterial Toxins, Division of Bacterial Products, Center for Biologics Evaluation and Research, Food and Drug Administration, Bethesda, MD 20892, [1]Department of Computer Science, The Pennsylvania State University, University Park, PA 16802, [2]Albert Einstein College of Medicine, New York City, NY 10461 and [3]National 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, 4.7×10^6 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 $\sim3\times10^9$ 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 anaylsis 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

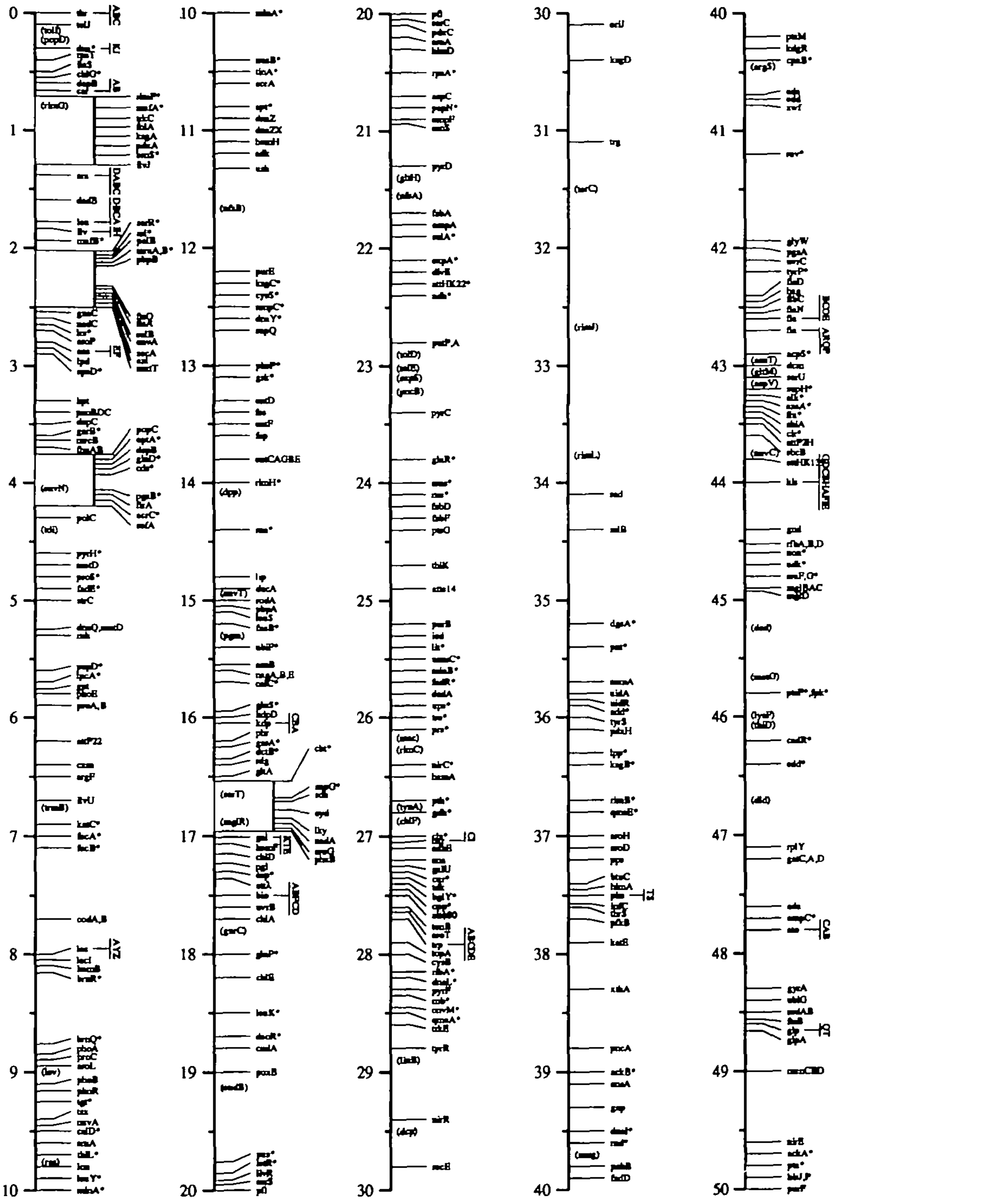

Figure 1. Computer drawing of a portion of the *E. coli* genetic map. The positions and gene names of the genes in the first 50 minutes of the 1983 *E. coli* genetic map (12) were computerized using the DigiMap program. The map was formatted using the Plasmid Description Language (see text) and printed on a laser printer.

adopted after experimenting with several alternatives, differs from the approaches of Churchill et al. (13) and Medigue et al. (14, see Discussion). Extensive evaluations (8) proved its superiority over a dynamic-programming approach for aligning sequence-derived restriction maps to our digital version of the *E. coli* genomic restriction map.

Our new version of MapSearch offers several improvements, while maintaining the same effectiveness. First, it implements an algorithm that is appreciably faster for long probes. Its execution time is now proportional to MP log P rather than to MP^2, where there are M map sites and P probe sites (W.M., Barr, J., and K.E.R., manuscript submitted for publication). Moreover, its versatility is greater as a result of a number of small additions. For example, 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.

[a]Gene	[b]Ori.	[c]Min.	[d]Rank	[e]*p* value	[f]Locus	[g]Acc. #	[h]Bp.	[i]Sites	[j]First	[j]Last	[k]DB	[l]Type
'dcC	-	68.300	1	<0.001	ECTDCRAB	X14430	6295	9	3159000	3165600	E	repeats
tdcC	-	68.300	2	<0.001	ECTDCRAB	X14430	6295	9	3172700	3179000	E	repeats
tdcC	-	68.300	3	<0.001	ECTDCRAB	X14430	6295	9	3329300	3335800	E	repeats
tdcC	-	68.300	4	<0.001	ECTDCRAB	X14430	6295	9	3145400	3152000	E	repeats
tdcA	-	68.300	10	0.996	tdcReco	ES8002	3000	3	3319700	3321800	ES	modified
IS5	-	ND	20	ND	INS5ECO	J01734	1300	2	3318100	3318800	G	repeats
fecA	ND	7.800	<100	ND	ECOFEC	M20981	2645	7	ND	ND	G	repeats
fecA	+	7.800#	1	0.806	ECOFEC	M20981	2645	7	4589800	4591500	G	repeats
fecB	-	7.800	8	1.000	M26397	M26397	4842	6	325400	328700	G	repeats
fecB	-	7.800#	1	<0.001	M26397	M26397	4852	6	4585500	4589800	G	repeats
fecB	-	7.800	6	1.000	fecEecoM	ES1033	7259	13	325400	333000	G	repeats
fecB	-	7.800#	1	<0.001	fecEecoM	ES1033	7259	13	4585500	4592700	G	repeats
iap	+	59.200	9	1.000	ECOIAP	M18270	1664	3	2890700	2891600	G	normal
iap	+	59.200	1	0.141	iap-eco	ES8003	1665	4	2890300	2891600	ES	modified
bioH	-	75.000	15	1.000	ECBIOH	X15587	2614	10	3613300	3615200	E	normal
bioH	-	75.000	1	0.011	bioHeco	ES3007	2537	4	3613300	3615200	ES	modified
purK	ND	12.200	1	0.287	ECOPUREK	M19657	2449	2	561700	562700	G	twosite
pin	+	25.800	1	1.000	ECOPINP	K03521	2614	2	1220400	1221700	G	twosite
dnaC	-	98.900	2	0.575	ECODNATC	J04030	2554	2	4678600	4679700	G	twosite
fic	-	74.200	58	1.000	ECOFIC1	M28363	2496	7	3560700	3562000	G	8enz
fic	-	74.200	1	0.158	ECOFIC1	M28363	2496	4	3560700	3562000	G	7enz
katG	+	89.150	4	0.966	ECOKATGA	M21516	2805	7	4212100	4213600	G	8enz
katG	+	89.150	1	0.191	ECOKATGA	M21516	2805	5	4212100	4213600	G	7enz
fhuE	-	15.950#	1	0.010	ECOFHUE	X17615	2900	4	1175300	1178500	E	normal
ptsG	+	24.400	1	0.008	ECOPTSG	J02618	1523	4	1174300	1175300	G	normal
ptsG	+	24.400	1	<0.001	ptsGecoM	ES1078	4124	7	1174300	1178500	ES	meld
sspG	-	69.950	5	ND	ssp-ecoM	ES1102	2507	2	3443200	3445500	ES	meld
hemA	+	26.700	2	0.185	hemAecoM	ES8001	3714	7	1275200	1277100	ES	meld
prs	-	26.100	1	0.001	ECOPRS	M13174	1785	6	1273600	1275200	G	normal
prs	+	26.100	1	<0.001	prs-ecoM	ES1077	5493	12	1273600	1277100	ES	meld
valS	-	96.800	1	<0.001	ECOVALS	X05891	3293	12	4556100	4559500	G	normal
pepA	-	96.500	1	0.035	ECXERB	X15130	2038	10	4559500	4561200	E	normal
valS	-	96.500	1	<0.001	valSecoM	ES1115	5325	21	4556100	4561200	ES	meld
gdhA	+	27.000#	43	ND	ECOGDHAK	K02499	1937	2	1861700	1862400	G	7enz
gdhA	+	27.000#	1	0.168	gdhAecoP	EM6002	6800	8	1857700	1863800	EM	Pmap
crp	+	73.500	33	ND	ECOCRP	J01598	1127	2	3556300	3556800	G	normal
crp	+	73.500	1	0.106	crp-ecoP	EM6001	10000	8	3554600	3564200	EM	Pmap
sodA	-	88.450	11	1.000	ECOSOD	X03951	1053	3	4179800	4180100	G	normal
sodA	+	88.450	1	0.001	sodAecoP	EM6003	7250	6	4176200	4183400	EM	Pmap

(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 that 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): twosite, probes have only two sites; 7enz, EcoRV information is ignored; Pmap, a published physical map is used as a probe. ND, no data.

the map location lies in one of those regions (see below). MapSearch output contains rank, score, orientation, physical map location, predicted genetic map location, standard deviation from the mean, and *p* value for any specified number of best alignments (Figure 4). An alphanumeric alignment is also given for as many 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

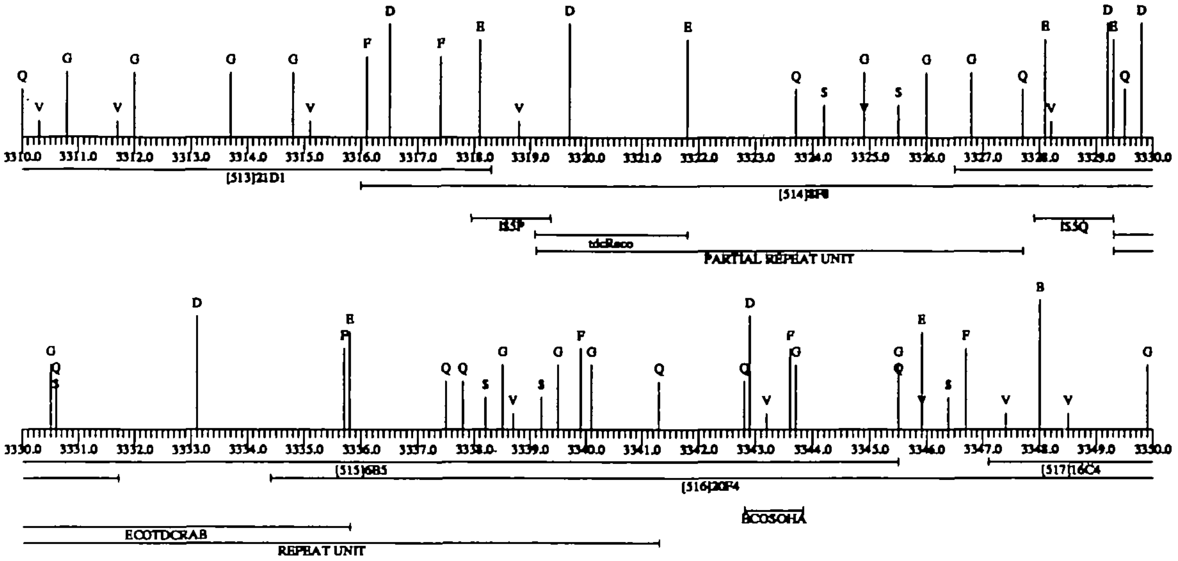

Figure 2. Restriction map of the *E. coli* chromosome (3310.0–3350.0 kilobases). PrintMap was used to depict the region containing the *tdc* operon. MapSearch was used to align the DNA sequences ECOTDCRAB and tdcReco (see text). The positions of the insertion elements IS5P and IS5Q, as well as the repeat unit and partial repeat unit data, are taken from Umeda and Ohtsubo (17). Three additional copies of the repeat unit are present elsewhere in the genomic restriction map (16,17, see Table 1). The position of the DNA sequence ECOSOHA (61) was determined using MapSearch (data not shown) and is consistent with published mapping data (61). Restriction enzyme sites and positions of the miniset clone inserts (513–516) were taken from Kohara et al. (2). Restriction enzymes: B, BamHI; D, HindIII; E, EcoRI; F, EcoRV; G, BglI; Q, KpnI; S PstI; V, PvuII.

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 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

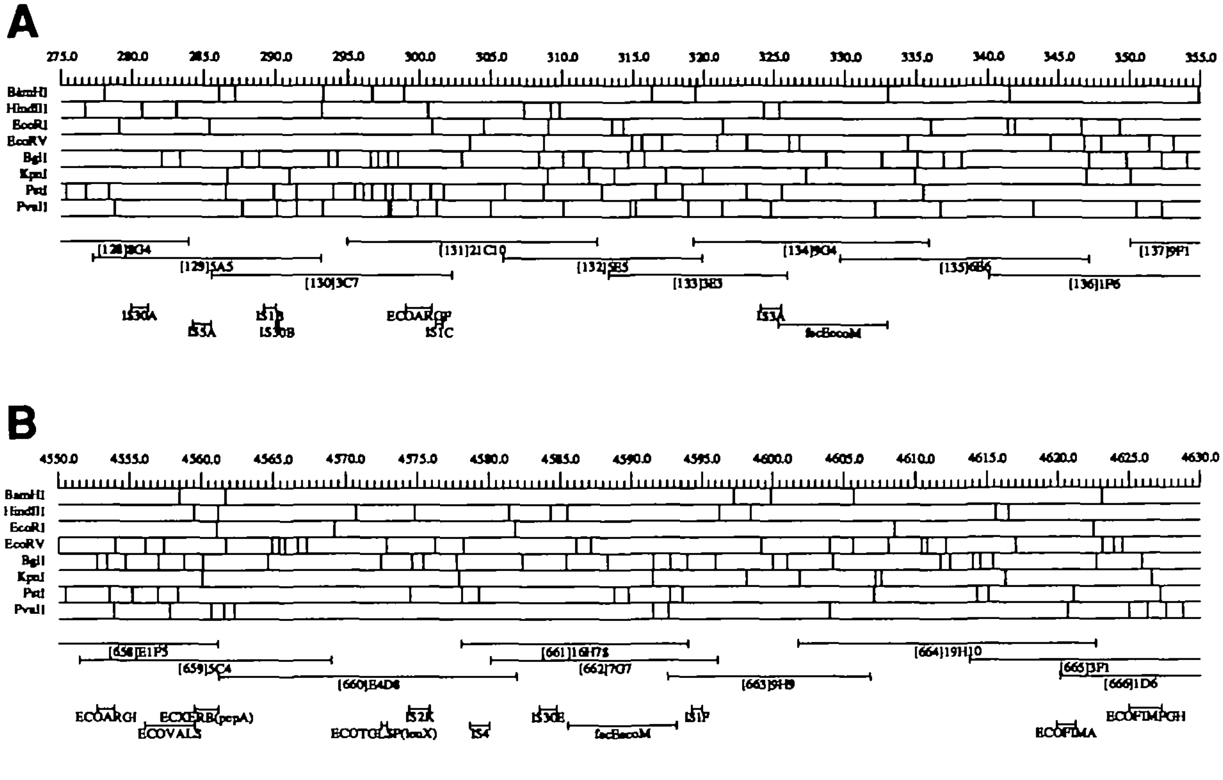

Figure 3. Restriction maps of the *E. coli* chromosome. A. The *E. coli* genomic restriction map from 275.0 to 355.0 kilobases is depicted. The *fec* operon (fecEcoM) location consistent with the genetic mapping data (1,21,22, see text) is in a region containing several insertion sequences (16,22,23) 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 *fec* operon (fecEecoM) is the highest-ranked MapSearch alignment (see Figure 5). The DNA sequences of the nearby genes *argI* (duplicate gene, ECOARGI, 63), *valS* (ECOVALS, 47), *pepA* (ECOXERB, 48), *leuX* (tRNA gene,ECOTGLSP,64), *fimA* (ECOFIMA, 65), and *fimFGH* (ECOFIMFGH, 66) were localized with MapSearch (Table 1, ref. 7, and data not shown).

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 dehyratase) 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 IS5 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 IS5 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 IS5 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 IS5 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 IS5. 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 IS5 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 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 IS5 and *tdcR* (the gene adjacent to *tdcA*) both contain a DNA sequence of 18/21 identical bp (data not shown).

```
A perfect score is 2600.
        score:  +/-     origin  st dev  prob
  1      2268    -      1175.5   3.55   0.010
  2      1541    +      4464.4   2.97   0.776
  3      1357    -      3766.9   2.83   0.995
  4      1298    +      3490.5   2.78   1.000
  5      1237    +       504.8   2.73   1.000
  6      1223    -      2002.5   2.72   1.000
  7      1202    +      1876.6   2.70   1.000
  8      1191    -      3529.9   2.70   1.000
  9      1102    +      3717.0   2.63   1.000
 10      1098    -       681.1   2.62   1.000
 11      1089    -       559.6   2.62   1.000
 12      1077    -      1245.0   2.61   1.000
 13      1027    +      3766.3   2.57   1.000
 14      1021    +       709.1   2.56   1.000
 15      1002    -      1387.5   2.55   1.000

        Alignment #1 (origin = 1175.5 kb):
    Left end in map at 1175295 bp.
Kpnl    2.895 kb <--> 1175.3 kb = 24.8 minutes
Pvu2    2.218 kb <--> 1176.2 kb = 24.9 minutes
Kpnl    0.607 kb <--> 1177.8 kb = 24.9 minutes
EcoR5   0.016 kb <--> 1178.5 kb = 24.9 minutes
    Right end in map at 1178516 bp.

        Alignment #2 (origin = 4464.4 kb):
    Left end in map at 4464184 bp.
EcoR5   0.016 kb <--> 4464.2 kb = 94.3 minutes
Kpnl    0.607 kb <--> 4465.2 kb = 94.4 minutes
Pvu2    2.218 kb <--> 4466.6 kb = 94.4 minutes
Kpnl    2.895 kb is unaligned
    Right end in map at 4467282 bp.
```

Figure 4. MapSearch output. The positions of the best 15 MapSearch alignments of the ECOFHUE DNA sequence on the genomic restriction map are given, as well as the best two alignments in alphanumeric format. The score and origin of an alignment are as in reference 8. Orientation and *p* value are as in Table 1. The standard deviation is relative to all possible alignments to the genomic map.

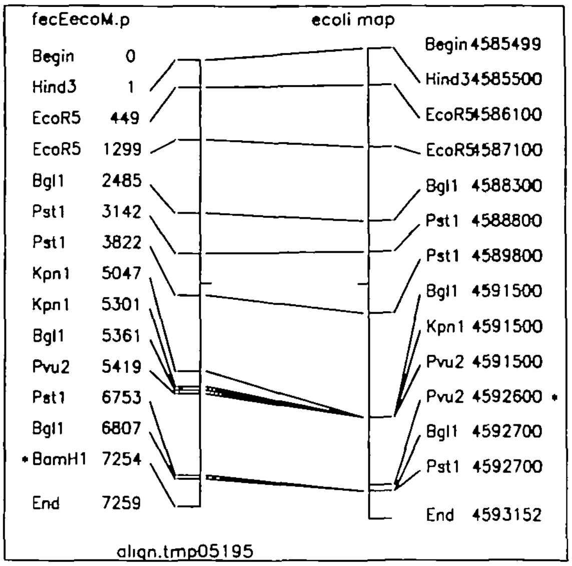

Figure 5. PrintAlign output. The highest-ranked alignment of the fecEecoM DNA sequence-generated restriction map to the genomic restriction map is depicted. This is not the location of the *fec* operon as determined by genetic experiments (see Table1 and text). Unaligned sites are denoted with '•'. Numbers denote DNA addresses in basepairs. The genomic map starts with 0 basepairs at the *thr* operon (2). The 4590000 region corresponds to 97 minutes on the genetic map (1).

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 *argI*/*argF* genes (1) (*argF* is approx. 30 kb. clockwise of the 7.8 minute locus and *argI* 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 iap-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 ECODNATC (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 in 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 ECOFIC1 (30) to the genomic restriction map only by ignoring the EcoRV sites in the probe. The ECOFIC1 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 ECOCPDB (32) provides an example of how genes can be misplaced due to the lack of EcoRV information in the genomic map. We previously placed ECOCPDB 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 ECOCPDB with a genomic PvuII site at 4516.1 kb (data not shown). The ECOCPDB 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 ECONARLX (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 ECOPTSG (39) and ECFHUE (40) individually yield significant highest-ranked alignments with MapSearch (Table 1). While the ECOPTSG alignment is consistent with the *ptsG* genetic map position (24.700 minutes), the ECOFHUE sequence was significantly aligned to only one position, far from the reported *fhuE* genetic map position of 15.950 minutes (1,41). The mapping of *fhuE* to 16 minutes was reported to have been difficult and the expected linkage to some genes in the 16 minute region (*tolA*, *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 *fhuE* 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 *rpsI* gene (70.250 minutes) sequence (ECORPSI, 43) are adjacent to one another. The first 223 base pairs of ECOSSPG are identical to bases 892 to 1114 of ECORPSI. However the last 70 bases of ECORPSI are not found in the ECOSSPG 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 *rpsI* gene is not present in most strains of *E. coli*, and its presence here has been previously noted, although the overlap with ECOSSPG was not previously detected (17). This DNA sequence meld, ssp-ecoM, 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 ECORF1X (45, containing the *prfA* 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

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 *valS* 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 ECOGDHAK (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 ECOCRP (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 already 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

(58) and IS sequences (17,23,24), we have placed sequences on the genomic restriction map that could not be aligned by any of our 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 DNA 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 view 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 electronic network 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'.

Note added

After this paper was accepted for publication the sequence of an IS1 element just upstream of the *E. coli fecA* operon was published in van Howe et al., (1990) J. Bacteriol., **172**, 6749–6758. A comparison with the DNA sequence of IS1F (Umeda M. and Ohtsubo E., ECIS1F, EMBL accession no. X52538, unpublished) confirms that the fecEecoM alignment depicted in Figure 3B and Figure 5 is the correct alignment.

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

1. Bachmann, B.J. (1990) Microbiol. Rev., **54**, 130–197.
2. Kohara, Y., Akiyama, K., and Isono, K. (1987) Cell, **50**, 495–508.
3. Kroger, M. (1990) Nucleic Acids Res., **18** (supplement), 2549–2587.
4. Church, G.M. and Kieffer-Higgins, S. (1988) Science, **240**, 185–188.
5. Daniels, D.L. and Blattner, F.R. (1987) Nature (London), **325**, 831–832.
6. Anderson, A. (1989) Nature (London), **338**, 283.
7. Rudd, K.E., Miller, W., Ostell, J. and Benson, D.A. (1990) Nucleic Acids Res., **18**, 313–321.
8. Miller, W., Ostell, J. and Rudd, K.E. (1990) CABIOS, **6**, 247–252.
9. Harman, D., Benson, D., Fitzpatrick, L., Huntzinger, R., and Goldstein, C. (1988) *In* Proceedings of RIAO 88 Conference 'User-Oriented Content-Based Text and Image Handling', pp. 840–848.
10. Williams, K M. (1988) CABIOS, 4, 211.
11. Pearson, W.R. and Lipman, D.J. (1988) Proc. Natl. Acad. Sci. U.S.A., 85, 2444–2448.
12. Bachmann, B.J. (1983) Microbiol. Rev., 47, 180–230.
13. Churchill, G.A., Daniels, D.L. and Waterman, M.S. (1990) Nucleic Acids Res., **18**, 589–597.
14. Medigue, C., Bouche, J.P., Henaut, A. and Danchin, A. (1990) Mol. Microbiol., **4**, 169–187.
15. Schweizer, H. P. and Datta, P. (1989) Nucleic Acids Res., **17**, 3994.
16. Muramatsu, S., Kato, M., Kohara, Y. and Mizuno, T. (1988) Mol. Gen. Genet., **214**, 433–438.
17. Umeda, M. and Ohtsubo, E. (1990) J Mol. Biol., **213**, 229–237.
18. Schweizer, H. P. and Datta, P. (1990) J. Bacteriol., **172**, 2825.
19. Schweizer, H. P. and Datta, P. (1988) J. Bacteriol., **170**, 5360–5363.
20. Pressler, U., Staudenmaier, H., Zimmerman, L. and Braun, V. (1988) J. Bacteriol., **170**, 2716–2724.
21. Staudenmaier, H., Van Hove, B., Yaraghi, Z. and Braun, V. (1989) J. Bacteriol., **171**, 2626–2633.
22. Zimmerman, L., Hantke, K. and Braun, V. (1984) J. Bacteriol., **159**, 271–277.
23. Umeda, M. and Ohtsubo, E. (1989) J. Mol. Biol., **208**, 601–614.
24. Birkenbihl, R.P. and Vielmetter, W. (1989) Mol. Gen. Genet., **220**, 147–153.
25. Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. and Nakata, A. (1987) J. Bacteriol., **169**, 5429–5433.
26. O'Regan, M., Gloeckler, R., Bernard, S., LeDoux, C., Ohsawa, I. and Lemoine Y. (1989) Nucleic Acids Res., **17**, 8004–8004.
27. Watanabe, W., Sampei, G., Aiba, A. and Mizobuchi, K. (1989) J. Bacteriol., **171**, 198–204.
28. Plasterk, R.H.A. and van de Putte, P. (1985) EMBO J., **4**, 237–242.
29. Masai, H., Bond, M.W. and Arai, K.-I. (1986) Proc. Natl. Acad. Sci. U.S.A., **83**, 1256–1260.

30. Kawamukai, M., Matsuda, H., Fujii, W., Utsumi, R. and Komano, T. (1989) J. Bacteriol., **71**, 4525–4529.
31. Triggs-Raine, B.L., Doble, B.W., Mulvey, M.R., Sorby, P.A. and Loewen, P.C. (1988) J. Bacteriol., **170**, 4415–4419.
32. Liu, J., Burns, D.M. and Beacham, I.R. (1986) J. Bacteriol., **165**, 1002–1010.
33. Rossi, J., Egan, J., Hudson, L. and Landy, A. (1981) Cell, **26**,305–314.
34. McCorkle, G.M. and Altman, S. (1982) J. Mol. Biol., **155**, 83–103.
35. Stewart, V., Parales, J.Jr. and Merkel, S.M. (1989) J. Bacteriol., **170**, 2229–2234.
36. Noji, S., Nohno, T., Saito, T., Taniguchi, S. (1989) FEBS Lett., **252**, 139–143.
37. Blasco, F., Iobbi, C., Giordano, G., Chippaux, M., Bonnefoy, V. (1989) Mol. Gen. Genet., **218**, 249–256.
38. Sodergren, E.J., DeMoss, J.A. (1988) J. Bacteriol., **170**, 1721–1729.
39. Erni, B. and Zanolari, B. (1986) J. Biol. Chem., **261**, 16398–16403.
40. Sauer, U., Hantke, K., Braun, V. (1990) Mol. Microbiol., **4**, 427–437.
41. Hantke, K. (1983) Mol. Gen. Genet., **191**, 301–306.
42. Serizawa, H. and Fukuda, R. (1987) Nucleic Acids Res., **15**, 1153–1163.
43. Isono, S., Thamm, S., Kitakawa, M. and Isono, K. (1985) Mol. Gen. Genet., **198**, 279–282.
44. Li, J.-M., Russell, C.S. and Cosloy, S.D. (1989) Gene, **82**, 209–217.
45. Craigen, W.J., Cook, R.G., Tate, W.P. and Caskey, C.T. (1985) Proc. Natl. Acad. Sci. U.S.A., **82**, 3616–3620.
46. Hove-Jensen, B., Harlow, K.W., King, C.J. and Switzer, R.L. (1986) J. Biol. Chem., **261**, 6765–6771.
47. Haertlein, M., Frank, R. and Madern, D. (1987) Nucleic Acids Res., **15**, 9081–9082.
48. Stirling, C J., Colloms, S., Collins, J.F., Szatmari, G., Sherratt, D.J. (1989) EMBO J., **8**, 1623–1627.
49. Heck, J.D. and Hatfield, G.W. (1988) J. Biol. Chem., **263**, 857–867.
50. Valle, F., Becerril, B., Chen, E., Seeburg, P., Heyneker, H. and Bolivar, F. (1984) Gene, **27**, 193–199.
51. McPherson, M.J. and Wootton, J.C. (1983) Nucleic Acids Res., **11**, 5257–5266.
52. Kim, S.Y., McLaggan, D. and Epstein, W. (1990) J. Bacteriol., **172**, 6127–6128.
53. Sanderson, K.E. and Roth, J.R. (1988) Microbiol. Rev., **52**, 485–532.
54. Riley, M. and Sanderson, K.E. (1990) *In* Drlica, K. and Riley, M. (ed.), The Bacterial Chromosome. ASM Press, Washington, D.C., pp. 85–95.
55. Aiba, H., Fujimoto, S. and Ozaki, N. (1982) Nucleic Acids Res., **10**, 1345–1361.
56. Takeda, Y. and Avila, H. (1986) Nucleic Acids Res., **14**, 4577–4589.
57. Touati, D. (1988) J. Bacteriol., **170**, 2511–2520.
58 Komine, Y., Adachi, T., Inokuchi, H. and Ozeki, H. (1990) J. Mol. Biol., **212**, 579–598.
61. Baird, L. and Georgopoulos, C. (1990) J. Bacteriol., **172**, 1587–1594.
62. Moore, S.K., Garvin, R.T. and James, E. (1981) Gene **16**, 119–132.
63. Bencini, D.A., Houghton, J.E., Hoover, T.A., Foltermann, K.F., Wild, J.R. and O'Donovan, G.A. (1983) Nucleic Acids Res. **11**, 8509–8518.
64. Thorbjarnardottir, S., Dingermann, T., Rafnar, T., Andresson, O.S., Soell, D. and Eggertsson, G. (1985) J. Bacteriol. **161**, 219–222.
65. Klemm, P. (1984) Eur. J. Biochem. **143**, 395–399.
66. Klemm, P. and Christiansen, G. (1987) Mol. Gen. Genet. **208**, 439–445