Alignment of *Escherichia coli* K12 DNA sequences to a genomic restriction map

Kenneth E.Rudd*, Webb Miller[1], James Ostell[2], and Dennis A.Benson[2]
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 and [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received September 22, 1989; Revised and Accepted December 7, 1989

ABSTRACT

We use the extensive published information describing the genome of *Escherichia coli* and new restriction map alignment software to align DNA sequence, genetic, and physical maps. Restriction map alignment software is used which considers restriction maps as strings analogous to DNA or protein sequences except that two values, enzyme name and DNA base address, are associated with each position on the string. The resulting alignments reveal a nearly linear relationship between the physical and genetic maps of the *E. coli* chromosome. Physical map comparisons with the 1976, 1980, and 1983 genetic maps demonstrate a better fit with the more recent maps. The results of these alignments are genomic kilobase coordinates, orientation and rank of the alignment that best fits the genetic data. A statistical measure based on extreme value distribution is applied to the alignments. Additional computer analyses allow us to estimate the accuracy of the published *E. coli* genomic restriction map, simulate rearrangements of the bacterial chromosome, and search for repetitive DNA. The procedures we used are general enough to be applicable to other genome mapping projects.

INTRODUCTION

The genome of *Escherichia coli* consists of a supercoiled, circular DNA molecule of 4.7 million base pairs (1). This single chromosome can be represented in three different ways:

1. As a genetic map, which is a series of genes that are identified by mutant phenotypes and are ordered using information from genetic crosses (2). Genetic maps portray linkages that aid in strain construction, and these maps are used to catalog information about genes by chromosomal locus.

2. As a physical map, which is a series of restriction endonuclease-generated DNA fragments ordered using molecular cloning, gel electrophoresis, and DNA hybridization techniques (1). Such maps aid the further cloning and sequencing of genes and are central to many molecular diagnostic procedures including Southern blotting, RNA mapping, and restriction site polymorphism typing.

3. As a DNA sequence. This last map of *E. coli* is, as of yet, incomplete, although more than 20% of the chromosome has been DNA-sequenced (3) and projects designed to produce a complete *E. coli* sequence are now underway (4). DNA sequencing permits precise length determination of any subregion of the chromosome and reveals signals encoded in DNA that define genes and operons.

Our ultimate goal is to integrate all three types of information into a single map, which we believe will have increased reliability due to checks for internal consistency. Integration will also lead to a refined genetic map because genes can be represented by the signals in their DNA sequence. In an integrated genomic map, the physical map can serve to align, order, and orient sequenced genes. Hence, a framework is provided for monitoring progress toward a completely sequenced genome. We have made progress towards forming an integrated genomic map using the 1983 *E. coli* genetic map (2), the genomic restriction map published by Kohara *et al.* (1), and DNA sequence database entries (see Table 1).

To improve access to the information contained in the *E. coli* genomic maps, we developed computer methods analogous to those used to align DNA and protein sequences (Miller *et al.*, submitted). In fact, our first map alignments were performed using a one-letter code for restriction enzymes and the FASTA protein sequence alignment program (5). This method ignores DNA fragment length information and was superceded by our new methods.

Many DNA sequences include genes whose genetic map positions are known. One can create links between the genetic and physical maps by calculating restriction maps from a DNA sequence (a sequence-derived restriction map is defined here as a probe), then finding the region of the genomic *E. coli* physical map with greatest similarity to a probe. We describe below the first steps toward the construction of a computerized integrated *E. coli* genomic map and make a number of observations based on this analysis of the physical and genetic maps.

* To whom correspondence should be addressed

Table 1. 199 GenBank (Release 59.0) and 2 EMBL (Release 18.0) *E. coli* K12 DNA sequence database entries were converted to restriction map probes and aligned to the genomic restriction map of Kohara *et al.* (1) using MAPSEARCH software. Only ECDUTPYR and ECEBGRA are from EMBL since we used the GenBank sequence when both databases had the same entry. Our partial list of sequences is derived from an unpublished listing kindly provided to us by G. Church. The 54 entries used to generate Figure 4 are marked with *. The 32 entries added after the original analysis are marked with #.

[a]Gene	[b]Ori.	[c]Pos. (min)	[d]Rank	[e]Addr. (kb)	[f]*p* value	[g]Locus	[h]Length (bp)	[i]Acc. num.
thrA*	(+)	0.000	1	00.1	0.002	ECOTHR	5922	J01706
dnaK #	(+)	0.300	1	12.7	0.013	ECODNAK	1917	K01298
dnaJ #	(+)	0.300	1	14.1	0.003	ECODNAJK	1623	M12544
rpsT	−	0.400	1	18.6	0.010	ECORPSTB	2882	X04382
ileS	+	0.500	2	20.7	0.810	ECORPSTA	1806	M10428
dapB		0.600	M			ECODAPB	1281	M10611
carA*	(+)	0.650	2	30.9	0.269	ECOCARAB	5227	J01597
folA		0.950	M			ECODHFOLG	1200	X05108
ksgA	(−)	1.025	4	50.5	0.808	ECOAPAH	2396	X04711
araB*	(−)	1.375	1	66.4	<0.001	ECOARAABD	4478	M15263
ilvI*	(+)	1.850	1	86.5	0.022	ECOILVIH	2323	X01609
pbpB	+	2.150	9	90.9	1.000	ECOPBPB	2759	K00137
ftsA*	(+)	2.350	8	102.6	0.966	ECOFTSQA	3333	K02668
ftsA		2.350	M			ECOFTSQAB	1870	M10429
secA	(+)	2.450	1	108.6	0.001	ECOSECA	3811	M20791
envA*	(−)+	2.450	5	108.8	0.994	ECOENVAA	2048	M19211
lpd*	(+)	2.850	1	123.8	<0.001	ECOACE	7740	V01498
fhuA	+	3.700	1	170.0	0.005	ECOFHUACD	4607	M12486
fhuB	+	3.700	1	172.7	0.583	ECOFHUB	2563	X04319
tsf'	(+)	4.000	1	201.0	0.873	ECORPSBTS	2192	J01684
dnaE #	+	4.500	1	211.6	<0.001	ECOLPXA	6627	M19334
rnh	−	5.300	1	246.0	0.009	ECORNHQ	1592	K00985
gpt		5.750	M			ECOGPTA	2253	M13422
phoE	−	5.800	1	268.5	0.809	ECOPHOE	1980	J01662
proA	+	5.900	1	269.4	0.003	ECOPHOEA	3041	X00786
argF*	(−)	6.500	1	299.9	0.614	ECOARGF	1405	X00759
lacZ*	(−)	8.050	1	370.1	<0.001	ECOLAC	7477	J01636
phoA*	(+)	8.825	1	410.2	0.016	ECOPHOAA	2715	M13345
proC		8.900	M			ECOPROC	968	J01665
phoB		9.100	M			ECOPHOB	976	X04026
phoR		9.150	M			ECOPHORG	1972	X04704
dnaZ	+	10.875	1	501.4	0.282	ECOZXPIII	2775	X04487
htpG #		11.100	M			ECOHSP	2235	M17218
ushA		11.325	M			ECOUSHA	1819	X03895
nmpC #		12.500	M			PA2LC	2816	J02580
fepA		13.625	M			ECOFEPAA	2624	M13748
dacA		14.875	M			ECODACA	1597	X06479
pbpA	−	15.050	1	681.1	0.126	ECOPBPA	2936	D00001
leuS	−	15.100	2	686.9	0.007	ECOLEUS	3618	X06331
rlpA	−	15.100	1	678.2	0.003	ECORLPA	1408	M18276
nagB*	+	15.600	1	717.2	0.050	ECONAGBE	3391	M19284
kdpA		16.050	M			ECOKDPABC	4933	K02670
phr		16.200	M			ECOPHRORF	2039	K01299
sucA*	+	16.750	1	763.4	<0.001	ECOGLTA	13063	J01619
cyd	+	16.775	1	783.6	0.143	ECOCYD	3845	J03939
galK*	(−)	17.000	9	799.6	1.000	ECOGALK	1622	X02306
bioA #	(+)	17.500	3	820.1	0.010	ECOBIO	5793	J04423
uvrB*	(+)	17.600	1	825.0	0.038	ECOUVRB2	2400	X03722
serS*	(−)+	19.950	1	950.6	<0.001	ECOSERS	1854	X05017
aroA		20.200	M			ECOAROA	1284	X00557
rpsA #	+	20.500	1	972.9	<0.001	ECORPSA	2412	J01682
pepN #	+	20.800	2	1001.7	0.018	ECOPEPN	3409	M15273
ompF		20.900	M			ECOOMPF	1808	J01655
pyrD*	+	21.300	1	1016.5	0.017	ECOPYRD	1357	X02826
ompA	−	21.800	1	1031.4	0.004	ECOOMPA	2270	J01654
divE		22.200	M			ECOTGS	1344	X00547
pyrC*	(−)	23.400	1	1135.4	0.018	ECOPYRC	2046	D00002
ptsG*	+	24.400	1	1174.3	0.062	ECOPTSG	1523	J02618
umuC #	+	25.500	1	1243.1	0.080	ECOUMUCD	2454	M10107
prs #	−	26.100	1	1273.6	<0.001	ECOPRS	1785	M13174
narC	(+)	27.050	2	1292.9	0.037	ECONARG	509	X01164
tyrT	(+)	27.150	6	1300.5	0.999	ECOTGY1	1949	K01197
trpA*	(−)	27.700	1	1329.9	<0.001	ECOTGP	7335	J01714
topA	+	27.900	1	1344.0	0.003	ECOTOPA	4071	X04475
cysB*	(+)	28.000	1	1347.2	0.221	ECOCYSB	1840	M15041
pyrF*	+	28.300	1	1357.4	0.004	ECOPYRF	1549	J02768
nirR	−	29.400	1	1413.9	0.012	ECONIRR	1641	J01608

pntA #	–	35.400	1	1692.9	<0.001	ECOPNTAB	3240	X04195
fumA #	–	35.700	5	1701.8	0.121	ECOFUMC	2250	X04065
fumA #		35.700	M			ECOFUMA	2409	X00522
manA	+	35.700	1	1705.5	0.633	ECOMANAA	1604	M15380
uidA		35.800	M			ECOUIDAA	2439	M14641
aroH		37.000	M			ECOAROH2	567	J01593
aroD		37.100	M			ECOAROD	1798	X04306
btuC		37.400	M			ECOBTUCED	2600	M14031
thrS	(–)	37.600	1	1811.9	<0.001	ECOTHRINF	7784	V00291
pabB		39.800	M			ECOPABB	1623	K02673
ptsL*	+	40.200	1	1921.4	<0.001	ECOPTSLPM	3188	J02699
cheR*	(–)	41.600	1	1977.2	<0.001	ECOCHE3	3063	M13463
tap*	(–)	41.600	5	1981.0	0.998	ECOCHE2	3465	J01705
cheA*	(–)	41.800	5	1983.5	0.998	ECOCHE1	1360	M13462
uvrC*	(–)	42.100	1	2005.8	0.074	ECOUVRC	4549	X03691
hag*	+	42.450	6	2013.2	0.996	ECOHAG	1667	M14358
flaA*	(+)	42.700	1	2030.3	0.027	ECOFLAA	1763	M12784
sbcB*	+	43.600	2	2092.6	0.223	ECOSBCB	1927	J02641
hisD		44.000	M			ECOHISD	1305	X03972
gnd		44.375	M			ECOGND	1887	K02072
mglB		44.875	M			ECOMGLB1	1384	X05646
metG #	+	45.700	1	2203.0	<0.001	ECOMETG	2346	K02671
dld #	+	46.700	1	2230.4	0.024	ECODLD	2340	X01067
ada		47.600	M			ECOADA	1324	M10211
nrdA*	+	48.500	1	2353.4	<0.001	ECONRDA	8554	K02672
glpA #	(+)	48.675	1	2364.6	<0.001	ECOGLPA	4739	M20938
purF	–	50.000	1	2443.0	0.368	ECOHISPUR	6172	J02800
hisT #	–	50.200	1	2449.5	<0.001	ECOHIST1	2323	X02743
gltX #	–	52.050	5	2533.0	0.159	ECOGLTX	1514	M13687
crr		52.250	M			ECOPTSHI	2850	J02796
purM	+	53.800	1	2621.7	0.024	ECOPURMN	2899	M13747
guaB*	(–)	53.900	1	2632.1	<0.001	ECOGUABA	3531	M10101
hisS		54.100	M			ECOHISS	1673	M11843
rnc	+	55.400	9	2706.5	1.000	ECORNC1	1076	X02673
aroF		56.700	M			ECOPHEAB	4509	M10431
rplS	–	56.750	1	2749.6	0.006	ECOTRMD	4586	X01818
recN #	+	57.500	1	2757.2	0.020	ECORECN	2224	Y00357
alaS*	(–)	58.200	3	2831.8	0.878	ECOALAS	2770	J01581
recA	(–)	58.250	15	2835.1	1.000	ECORECA	1390	J01672
iap #		59.100	M			ECOIAP	1664	M18270
pyrG	–	59.700	1	2922.7	0.017	ECOPYRG	2442	M12843
relA* #	–	59.800	1	2925.9	0.005	ECORELA	2858	J04039
argA		60.500	M			ECOARGA	1575	Y00492
recB*	–	60.600	1	2966.8	<0.001	ECORECB	3960	X04581
recB		60.600	M			ECORECD	2160	X04582
recC	–	60.650	1	2973.1	<0.001	ECORECC	6000	X03966
lysA*	+	61.400	1	2990.2	0.010	ECOGALLYS	4295	J01614
araE #	–	61.300	1	2994.1	<0.001	ECOARAEA	2866	J03732
serA*	–	62.800	11	3070.3	0.999	ECOSERA	1233	N00029
metK*	+	63.700	1	3100.3	0.952	ECOMETK	1462	K02129
metC*	+	65.000	1	3205.9	0.298	ECOMETC	1880	M12858
cca*	+	66.800	1	3257.1	0.121	ECOCCA	2257	M12788
rpsU	(+)	67.000	1	3261.6	0.005	ECORPSU	4644	M16194
rpoD	(+)	67.000	1	3266.9	0.001	ECORPSRPO	5059	J01687
ebgR	+	67.800	1	3278.5	<0.001	ECEBGRA	4265	X03228
pnp	(–)	68.825	1	3375.4	0.055	ECORPSOP	3030	J02638
infB	–	68.900	1	3379.4	0.002	ECONUSA	5423	X00513
gltB	+	69.400	1	3420.9	<0.001	ECOGLTB	6292	M18747
rplQ	(–)+	72.400	15	3502.4	1.000	ECORPA	3154	X02543
rpsM		72.550	M			ECORPLP2	759	M12432
rplO		72.600	M			ECORPLN	5922	X01563
rpsJ	(–)	73.250	1	3519.6	<0.001	ECORPOS10	5422	X02613
tufA	(–)	73.325	12	3540.9	1.000	ECOSTR3	1374	J01690
fusA		73.325	M			ECOSTRA	2076	X00415
rpsL	(–)+	73.400	13	3544.9	1.000	ECOSTR1	1016	J01688
dam		74.350	M			ECODAM	1134	J01600
ompR	–	74.800	2	3605.8	0.704	ECOOMPB	2703	J01656
malT	(+)	75.200	5	3624.3	0.962	ECOMALT	3508	M13585
malP	(–)	75.200	3	3629.2	0.971	ECOMALP2	2600	X06791
glgA		75.400	M			ECOGLGA	1601	J02616
glgC	–	75.400	1	3640.0	0.027	ECOGLGC	1328	J01616
glgB	–	75.400	1	3643.1	0.002	ECOGLGBA	2559	M13751
asd		75.500	M			ECOASD	1674	V00262
livJ		75.900	M			ECOLIVJK1	1101	M10426

htpR #	–	76.400	1	3671.1	0.024	ECOHTPRR	1312	K02178
glyS #	–	79.500	6	3794.7	0.134	ECOGLYS	3333	J01622
xylA	–	79.700	1	3799.8	0.249	ECOXYLABA	4176	X00772
pyrE	+	81.800	2	3886.5	0.390	ECDUTPYR	2568	V01578
uhpT		82.100	M			ECOUHP	5400	M17102
ilvB #	–	82.200	1	3923.1	0.050	ECOILVBPR	2470	J01633
gyrB*	–	82.950	1	3949.3	0.004	ECORECFA	4931	X04341
dnaA*	(–)	83.050	1	3954.6	0.019	ECODNAAOP	3873	J01602
bglC #	(–)	83.450	1	3979.8	<0.001	ECOBGLO	5270	M16487
phoS		83.600	M			ECOPHOS	5032	K01992
uncI*	(–)	83.875	1	3989.6	<0.001	ECOUNCC	14526	X01631
asnA	+	84.000	1	4003.5	<0.001	ECOORIASN	4012	K00826
rbsK #	+	84.325	1	4011.5	0.003	ECORBS	5820	M13169
ilvG*	(+)	84.600	1	4028.6	<0.001	ECOILVGE	9456	M10313
rep*	+	84.700	1	4037.3	0.012	ECOREPHEL	2671	X04794
rho		84.750	M			ECORHO	1880	J01673
hemC #	–	85.300	1	4067.4	0.089	ECOHEMC	1957	X04242
cyaA*	+	85.000	1	4068.3	0.003	ECOCYAG	3699	K02969
uvrD	(+)	85.150	1	4075.3	0.010	ECOUVRD02	2846	X04037
uvrD*	(+)	85.150	1	4076.1	0.002	ECOUVRD	2869	X00738
pldA #	+	85.400	2	4082.4	0.019	ECOPLDAA	1319	X02143
polA	(+)	86.600	1	4125.5	0.416	ECOPOLA	4127	J01663
glnA*	(–)	86.700	1	4132.6	0.645	ECOGLN	4311	X05173
tpiA		88.250	M			ECOTPIA	1338	X00617
cdh	+	88.350	1	4186.2	0.030	ECOCDHA	3304	X02519
glpK	–	88.400	2	4194.8	0.290	ECOGLYK	2028	M18393
cytR*	–	88.800	1	4202.3	0.943	ECOCYTR	1384	X03683
metL	+	88.000	12	4208.2	1.000	ECOMETL	2433	J01651
metF		88.000	M			ECOMETF	1238	V01502
ppc*	–	89.450	1	4228.7	0.008	ECOPPCG	3106	X05903
btuB	+	89.600	1	4241.8	0.001	ECOBTUB	2220	M10112
birA #		89.650	M			ECOBIRA	2491	M10123
rrnB	+	89.750	1	4243.8	<0.001	ECORGNB	7508	J01695
tufB		89.800	M			ECOTGTUFB	1973	J01717
rpoB	(+)	90.000	1	4256.7	<0.001	ECORPLRPO	12337	J01678
aceK		90.600	M			ECOICDHKP	2214	M18974
lysC		91.200	M			ECOLYSCP	645	X00008
lysC*	(–)	91.200	1	4310.9	<0.001	ECOLYSC	1587	M11812
malG	(–)	91.500	1	4319.4	0.002	ECOXYLE	2842	J02812
lamB*	(+)	91.500	1	4323.4	0.002	ECOMALB	6545	J01648
plsB	–	91.900	1	4332.5	0.003	ECOPLSB	3865	K00127
tyrB	+	91.950	3	4347.2	0.754	ECOTYRBA	1733	M12047
uvrA*	(–)	92.000	1	4350.8	0.001	ECOUVRAA	3205	M13495
fdhF #	–	92.500	5	4377.2	0.099	ECOFDHF	2273	M13563
melA	+	93.400	1	4421.5	0.377	ECOMELOPA	1628	M18425
melB	+	93.400	1	4424.3	0.005	ECOMELB	1575	K01991
aspA	–	94.100	1	4446.6	0.001	ECOASPAG	2921	X04066
ampC	(–)	94.300	1	4458.3	0.002	ECOAMPCFR	5482	J01611
psd	–	94.600	1	4469.8	0.022	ECOPSD	1350	J03916
rpsR #	+	95.500	1	4505.4	0.002	ECORPSFRI	1979	X04022
cpdB	–	95.700	6	4526.9	0.988	ECOCPDB	2198	M13464
pyrB		96.500	M			ECOPYRBI	1593	J01670
argI		96.600	M			ECOARGI	1085	X00210
valS*	–	96.800	1	4556.1	0.018	ECOVALS	3293	X05891
pilA		98.000	M			ECOFIMA	1450	X00981
pilA		98.000	M			ECOPAPA	2110	X03391
pilC		98.000	M			ECOPAPC	2929	Y00529
hsdSK*	–	98.500	1	4657.1	<0.001	ECOHSDSK	2528	J01632
phoM	+	99.825	1	4712.5	<0.001	ECOPHOM	4658	M13608
dye	–	99.900	3	4716.4	0.721	ECODYE	1468	M10044

(a) Gene aligned to genomic restriction map. If the sequence entry encompassed several mapped genes, one was chosen to identify the entire sequence. (b) Orientations of the aligned genes as derived with 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. Orientations with parentheses are identical to those given in Bachmann et al. (2) and occasionally differ from our result. (c) The map position (in minutes) of the sequenced gene. The positions are approximated from the 1983 *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. The late entries (#) have minutes that were in some cases taken from GenBank, not the 1983 genetic map. (d) The rank of the alignment that best fits both physical and genetic map data. M is listed if none of the top fifteen alignments is commensurate with the genetic map position. These misses were excluded from Figure 3. (e) Genomic address coordinates (kb) of the first genomic restriction sites that are aligned to restriction sites in the DNA sequence probes. These are positions of the entire sequence entry, not neccessarily the particular gene chosen to identify the sequence entry. In order to align genetic and physical map data it was necessary to simulate reversion of the IN(*rrnD-rrnE*)*1* genome rearrangement. Coordinates that lie within the IN(*rrnD-rrnE*)*1* inversion can be easily converted to the published map coordinates (see text). (f) Probability value for each alignment (see text). (g) The name of the database entry.(h) The length in base pairs of the DNA sequences converted to probe maps. (i) Database accession numbers of files which contain detailed information used in this analysis.

RESTRICTION MAP ALIGNMENT SOFTWARE

We began by digitizing an *E. coli* genomic physical map (1) and the 1983 genetic map (2) to transform them from the graphical representations available in the published accounts to forms suitable for manipulation by computer. Each of the eight lanes of an enlarged copy of the genomic restriction map (1) [kindly provided by Y. Kohara] was converted separately to digital format using a tablet digitizer. These computer files were then combined into a single file containing chromosome position and enzyme name information. The relative order of all enzyme sites was also separately recorded with a digitizer ignoring address information, thus creating a string of sites (e.g. EBDVSS). The resulting restriction enzyme string map (abbreviated as RESM) was compared to the digital restriction map to identify missed sites and position errors large enough to cause changes in relative order. Kilobase addresses were obtained by adding the restriction fragment lengths and calibrating the total digitizer units to a total of 4719.6 kilobases, a value obtained by inspection of the published map. Several gaps in the restriction map were arbitrarily set at 2 kilobases (1) and we have done the same. For these reasons, the digital kilobase coordinates correspond closely to those published by Kohara *et al.* (1). For the few sites that were ambiguous our digital map represents an interpretation of the published data. We identified 7112 restriction enzyme sites. 489 map positions have two enzyme sites at the same address and 23 addresses have three enzyme sites. This is due to the fact that we have rounded the digitized coordinates to the nearest 100 basepairs.

An initial panel of 169 DNA sequences was selected from the GenBank® and EMBL DNA sequence databases. For this step we used only sequences containing: (a) three or more sites recognized by the eight restriction enzymes (BamHI, HindIII, EcoRI, EcoRV, BglI, KpnI, PstI, PvuII) used to make the genomic restriction map, and (b) the coding region for a gene that is positioned accurately on the 1983 *E. coli* genetic map (see Table 1). Probes were generated from these mapped gene sequences by computing the recognition sites of the restriction enzymes used to make the genomic restriction map, using the Mount-Conrad-Myers Sequence Analysis Software Package (6).

Software that efficiently aligns such calculated probes with experimentally determined restriction maps requires us to define restriction map alignment precisely and assign a penalty for deviation from perfect matches. Waterman *et al.* (7) defined restriction map alignments as shown in Fig. 1, panel A:

1. Alignments are assigned scores on the basis of strict one-to-one correspondence of ordered sites. Either the D sites or the C sites can be aligned. Alignment of all three E sites is not allowed.

2. Sites not paired with another site are penalized as misalignments, e.g. sites D, E, and F in map I and site D in map II.

3. Even if sites are paired correctly, penalties are assessed for discrepancies noted when comparing distances between adjacent sites.

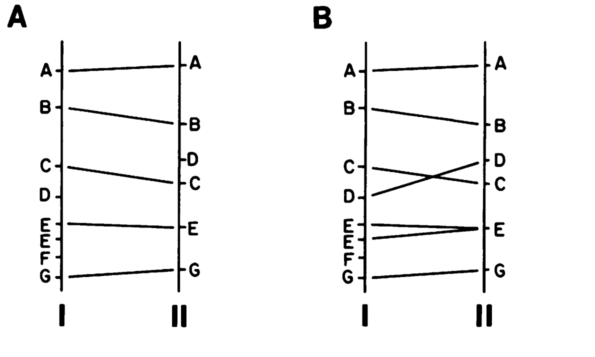

Figure 1. Alternative definitions of the alignment of two restriction maps, I and II. A. As defined by Waterman *et al.*(7), this alignment of restriction maps I and II has failed to align restriction sites D, E, and F in map I and D in map II. Penalties are assessed for all of these unaligned sites. B. A redefinition of restriction map alignments adopted in the algorithm used in this study would leave F as the only unaligned site.

Experiments with this approach and several others (8–10, Miller *et al.*, submitted) led us to redefine an alignment (Figure 1, panel B) to account for errors commonly produced during restriction mapping based on gel electrophoresis. Namely, two closely spaced sites for different restriction enzymes can be mistakenly reported in inverted order, and two proximal sites for the same restriction enzyme are often incorrectly reported as a single site. Our definition of alignment allowed alignment lines to cross (e.g., the C and D sites). We also allowed a site on the restriction map to align to two (or more) sites on the probe (e.g. the E sites on map I). Penalties for fragment length deviations and misaligned sites were combined by defining a linear relation between them. We have empirically determined that setting the penalty for one misaligned site equivalent to the penalty for any value between 400 and 1000 basepairs of distance discrepancy produces nearly equivalent end results (Miller *et al.*, submitted).

Our first step beyond the use of the FASTA program was to adapt a dynamic programming algorithm of Waterman *et al.* (7) to our problem. As originally stated, the method requires order M^2P^2 time, where there are M map sites and P probe sites. Algorithm improvements have reduced the time to order MP log P (X. Huang and E. Myers, personal communication). Unfortunately, we could not accommodate all features of our redefinition of alignment with these dynamic programming methods. Therefore we developed a non-dynamic algorithm, MAPSEARCH, which runs in order MP^2 time (Miller *et al.*, submitted). Using MAPSEARCH, we aligned the probes to the genomic restriction map. To estimate the shape of the score distribution for the MAPSEARCH algorithm, a sample of the 169 probes was run against 1000 shuffles of the map (Miller *et al.*, submitted) and the score of the best alignment noted. For probes with more than four sites, the distributions of best scores strongly resembles the extreme value distribution (11), whose cumulative distribution function, f(x), is

$$e^{-e^{-\lambda(x-u)}}$$

We estimated the parameters λ and u, which describe the distribution of maximal scores for a given probe, by $\lambda = \pi/\sqrt{6V}$ and $u = \mu-\gamma/\lambda$ where γ is Euler's constant (0.577...), μ is the sample mean, and V is the sample variance (12–15). The extreme value distribution has been shown to hold for certain simple scoring schemes, although there is some tendency to underestimate probability values (13,14). We established experimentally that the λ and u calculated from 100 shuffles of the map closely approximate those calculated from 1000 shuffles (Miller *et al.*, submitted). With 100 shuffles, it was quite feasible to compute λ and u for each of the 201 probes ultimately used for this analysis. The probability, *p*, of that probe producing a best alignment of score at least s is given by the formula

$$p = 1.0-e^{-e^{-\lambda(s-u)}}$$

COMPARISON OF GENETIC AND PHYSICAL MAPS

When we aligned the original 169 probes to the genomic restriction map, the highest scoring alignments for 75 probes were found to have a significance higher than 95% ($p < 0.05$). When the kilobase location of a probe, determined solely by computer alignment to the physical map, was plotted against the reported genetic position of a gene within it, a nearly linear relationship between the genetic and physical maps became clearly evident (Figure 2). The 'endpoints' of the linear maps represent the same point on the circular chromosome, defined as the *thr* locus (1–3). The six points lying off the line represent probes where the best scoring alignments were far from their genetic location and the proper genetic location was identified as a lower-ranking alignment, as was the case for numerous other probes (see Table 1).

Once the approximate relationship between the genetic and physical maps was established, we subjected the original 169 probes and an additional 32 probes to further analysis. These additional probes either contained genes that were less accurately mapped than those in the first 169 probes or were derived from sequences that became available after we completed our primary analysis. We developed the following method for aligning genes to the physical map. We first 'pinned' regions of the map by selecting all probe alignments with significance values of better than 95% and with reported genetic positions within 3.5 minutes of the genetic position calculated from their physical map position (based on the linear relationship mentioned above). Regions between pins were then filled in with alignments of less than 95% significance, but which were consistent with the position estimated from the flanking pins and which were among the top 15 alignments for that probe. When alternate possible alignments for a probe fell in the same interval, the best fit was selected. When an alignment placed by this method was found to be out of gene order with respect to more significant alignments flanking it, it was removed. This process enabled us to place 147 of the 201 probes (73%) onto the map. These alignments are summarized in Table 1.

A linear regression (16) through these 147 points indicates that:

$$\text{Minutes} = (\text{Kilobases} \times 0.021) - 0.001$$

Assuming a length of 4,719.6 kb (1) and 100 minutes (2) for the *E. coli* chromosome, the ideal line is drawn from (0,0) to (4719.6,100), represented by:

$$\text{Minutes} = (\text{Kilobases} \times 0.02119)$$

The ideal line differs very little from the regression line. However, when the differences between the actual minute positions and those predicted using the ideal linear relationship (the residuals) are plotted, a striking pattern emerges (Fig. 3). Genetic map positions were consistently underestimated (as compared to those predicted by the ideal line) in the interval from approximately 40 to 80 minutes (1900 to 3800 kb). This effect can even be seen with the data in Figure 2.

The question arises whether this small but consistent deviation from strict linearity has biological significance or is an artifact of data processing. Gene positions are often changed in new editions of the genetic map in an attempt to improve the interpretation of genetic data into positions on the chromosome (17,18). The residuals of Figure 3 show a clustering pattern because many genes are positioned relative both to neighboring genes and to 52 well-mapped genes that have been placed on a reference map (17). Thus some error is spread to adjacent sequences.

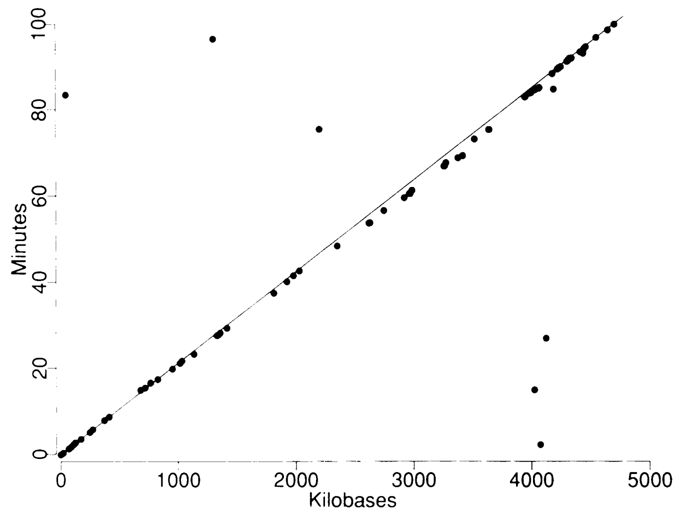

Figure 2. The relationship between physical and genetic map locations. 75 out of 169 DNA sequence-generated restriction maps (probes) had probability values <0.05 for the highest ranked MAPSEARCH alignments. The kilobase coordinate of the first aligned restriction site for each of the probes is plotted against its 1983 genetic map position (minute) demonstrating a nearly linear relationship. The ideal relationship is depicted by a straight line from the origin to (4719.6, 100). A best fit curve was generated using the UNIX S statistics package (16).

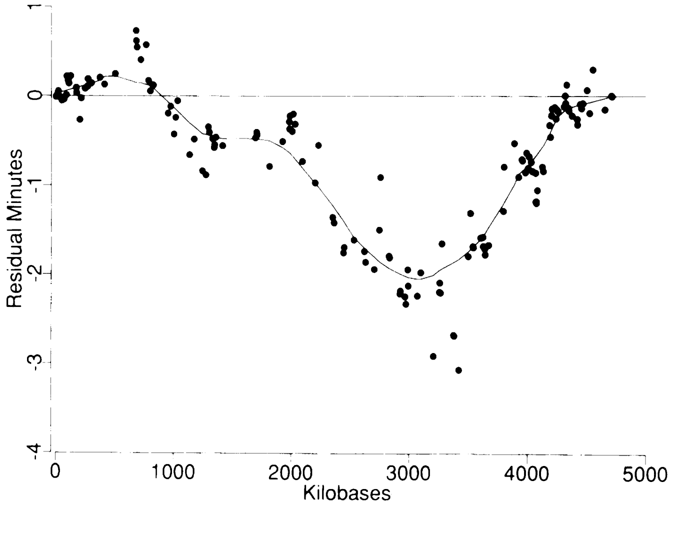

Figure 3. Variations of the published genetic location from the predicted placement for 147 DNA sequences. The difference between the published minute value and the value predicted from the physical map alignments (the residual minutes) is plotted against physical map position.

Fifty-four of our successfully aligned genes were also in the 1976 and 1980 genetic maps (see Table 1). The residual minutes were calculated using the earlier map positions and plotted as in Figure 3. The best fit residual curves for the three maps are shown superimposed on each other in Figure 4. The periodic changes made to improve the genetic map have consistently moved the gene positions closer to those predicted by the ideal linear relationship and physical map gene alignments. This improvement can be seen in the shrinking residuals of Figure

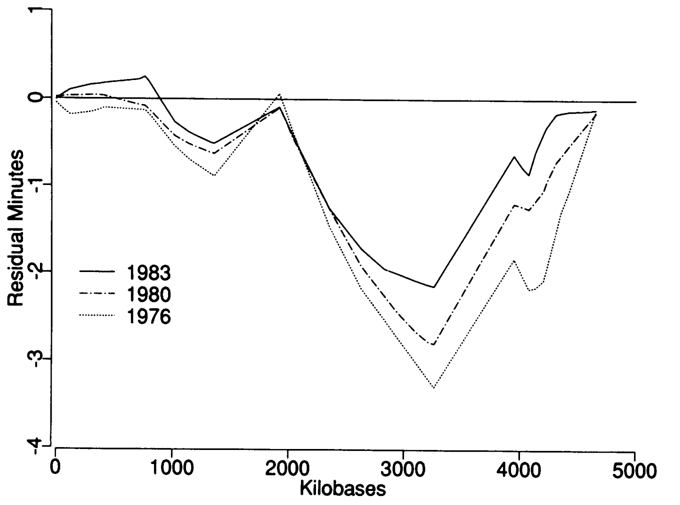

Figure 4. Comparison of residual minutes using the 1976, 1980, and 1983 genetic map positions (2,17,18). 54 genes had accurately placed positions on all three genetic maps. The residuals generated using these three different maps as compared to the ideal line are represented by the best fit curves as in Figure 3.

4. This suggests that the movement with improved genetic maps would be completed by producing a map, measured from 0 to 100 map units, based on the physical location of the genes on a genomic restriction map. The minutes of the *E. coli* genetic map are generally defined as the amount of DNA transferred during one minute of mating. However, a value of 45 kb/minute was used for those portions of the 1983 map that have physical map data linking genes (2). Since the *E. coli* genomic restriction map links genes around the entire 100 minute chromosome and we estimate the entire length to be 4719.6 kb, redefining a minute as 47.2 kb seems appropriate. This is preferred over a mating time standard because DNA transfer rates during mating can vary 15% or more due to strain differences and variations in culture conditions (17). The minutes of the closely related *Salmonella typhimurium* genetic map have already been defined as 45 kb (19).

ACCURACY OF THE PHYSICAL MAP

We can estimate the overall accuracy of the genomic physical map, which usually produces a less reliable restriction map than does DNA sequence. Discrepancies between probes and the aligned portion of the genomic restriction map occurred even with highly significant alignments and could reflect strain differences, inaccuracies in the physical map, or errors in the DNA sequences themselves. Examining the alignments reported in Table 1, we observe that about 10% of the restriction sites within aligned fragments could not be paired, even using our redefinition of alignment. The mean length differences between 488 corresponding pairs of single enzyme restriction fragments was 14.7% (which equals the sum of the absolute values of the difference between sequence derived and restriction map derived paired fragment lengths divided by the sum of the sequence derived fragment lengths).

We can also assess the accuracy of the genomic restriction map relative to the sequence database by looking at the sequences we did not assign to a genomic location (see Table 1). Most of the nonaligned sequences (denoted 'M') are not very information-rich (i.e., have relatively few restriction sites). However, the ECORPLN (72′) sequence contains 14 enzyme sites in 5922 bp and its best fit is more than two minutes away from its predicted position. It should be adjacent to the ECORPA (72′) sequence, another cluster of ribosomal protein genes. Although placed on the map at 3502.4 kb, ECORPA (72′) is also a poor fit and is in an orientation opposite to that reported by Bachmann *et al.* (2, Table 1). Both of these genes map near the rRNA gene *rrnD*. This is one of two inversion points that bracket the IN(*rrnD-rrnE*)*1* inversion present in the *E. coli* K12 W3110 strain whose genome has been restriction mapped (1). In our version of the genomic restriction map, the reinversion to wildtype has been simulated. The IN(*rrnD-rrnE*)*1* inversion endpoints were determined using the *rrn* operon alignments for *rrnD* and *rrnE* (see below) and crossover points known to be somewhere within the homologous 23S genes (20; C. Hill, personal communication). The genomic restriction map we used is inverted between 3493.0 and 4290.5 kb with respect to the map of Kohara *et al.* (1), thus resembling most *E. coli* K12 strains. These inversion points are arbitrarily chosen points within PvuII-EcoRI fragments at *rrnD* and *rrnE* that lie wholly within the 23S genes. (Kilobase coordinates, such as those given in Table 1, that lie within the inverted segment can be interconverted between our values and those of Kohara *et al.* by subtracting them from the sum of the two inversion points, 7783.5 kb.)

It is possible that our inability to align DNA sequence to this region of the map is due to the presence of additional chromosomal rearrangements in strain W3110. The IN(*rrnD-rrnE*)*1* inversion may be indicative of chromosomal instability. In support of this idea, Tabata *et al.* (21) have recently reported that their ordered cosmid genomic library of W3110 DNA provides evidence for a spontaneous DNA translocation relative to the W3110 strain that Kohara *et al.* (1) restriction mapped.

Another sequence that we did not place is PA2LC (12′), part of the PA2 prophage containing the *nmpC* gene, located between *dnaZX* at 10.9 min and *rlpA* at 15.1 min. We found two reasonable fits to PA2LC (12′) in inverted orientations in this interval, so we could not choose either one. In another instance, ECOGLGA (75′), we rejected an alignment because it fell inside another sequence and we determined there was no DNA homology. Finally, in some cases, e.g. ECOUHP (82′), pinning probes of low information is made difficult because regions of the original genomic restriction map lack information for sites of one of the restriction enzymes (EcoRV). In a number of cases, small probes can be pinned to the map by utilizing additional information, such as the restriction map of the parent clone of the sequenced DNA, but this was not done for this study.

SEARCH FOR REPEATED DNA

The repeated nature of a number of regions of the genomic restriction map became evident during our work. Genes placed on the map at better than 95% confidence, but not using the top alignment (see Table 1) can have (presumably incorrect) alignments to other places on the map that are more significant, e.g. ECOLEUS (15′), ECOBIO (17′), ECOPEPN (20′), ECONARG (27′), and ECOPLDAA (85′). We do not know if these restriction map similarities reflect homologous DNA segments, although DNA sequence or hybridization analysis could be used to resolve this point. In general, a modest divergence in DNA sequence produces a large map difference. For example, the *Salmonella typhimurium ara* operon DNA sequence (22–24) is 83.7% identical to its *E. coli* counterpart (25), yet shares only

4 of its 11 probe enzyme sites with the *E. coli ara* probe. We were unable to align the *Salmonella ara* sequence to the *E. coli* genomic restriction map at the proper genetic location. We have begun a search for large restriction map repeats. The most obvious repeating pattern had four copies, three of which were in tandem array. This repeating map pattern was found associated with IS5 insertion elements by Muramatsu *et al.* (26) and we detected it near 3200 kb during a string search for self similarities of restriction map sites. An 8.3 kb palindromic DNA restriction map pattern was located at 1648.0 kb. We were able to locate all seven known copies of the rRNA operons using the *rrnB* (ECORGNB) DNA sequence and genetic map information. We used the map locations of the *rrn* operons (2,27) and the ECORGNB (89′) sequence (see Table 1) to locate the EcoRI site in the 16S gene of the *rrn* operons at the following kb addresses (addresses inside the IN(*rrnD-rrnE*)*1* inversion can be converted to W3110 map coordinates as described above): *rrnA*, 4114.4; *rrnB*, 4245.7; *rrnC*, 4020.5; *rrnD*, 3495.8; *rrnE*, 4287.5; *rrnG*, 2735.2; *rrnH*, 246.0. Similarly, the multiple copies of another repeat sequence, *rhs* (28,29), were located by alignment searches (data not shown).

AN INTEGRATED GENOMIC MAP

Ultimately, we were able to align 466.2 kb of DNA sequence (9.9%) to the genomic restriction map, or about 50% of the *E. coli* sequence data present in the DNA databases (3). The majority of the remaining reported sequences reside in small segments (fewer than three restriction enzyme sites) that prevent accurate alignment. 110 out of 147 (68%) of our aligned probes are assigned their highest scoring alignment. Most of these alignments score substantially higher than the second-best alignment, which raises our confidence in the reported match. Table 1 summarizes the final results of the entire alignment process for each of the 201 probes analyzed. The alignment coordinates we report represent the best fits to the combined genomic physical and genetic map data. Three classes of sequence alignment are reported: a) sequences pinned to the map because of highly significant top rankings (see Figure 2), (b) less significant alignments made utilizing genetic map locations (as in Figure 3), and (c) sequences which were not aligned by our procedure (designated by M). We identify the gene whose map location was used to align a particular DNA sequence. Database entries are identified by their easily remembered locus name and by a permanent accession number. Using these accession numbers one can obtain the literature citations for the original *E. coli* DNA sequence data either from the databases directly or from a recent review article (3).

As a byproduct of the alignment process, we obtained information about the probable orientation of gene transcription with respect to the genomic map. The genomic map coordinates of aligned sequences are given, positioning the DNA sequences and the genes contained within them on the genomic map. Table 1 also shows the MAPSEARCH ranking and probability estimate for each of the alignments. These allow one to assess the ability of the MAPSEARCH software to align probes as a function of map location and sequence length. The *p* value is an indication of how uniquely and how well restriction map patterns match. Our confidence in a position assignment is bolstered by a highly significant alignment, however we cannot predict or preclude a homologous relationship between DNA segments on the basis of *p* value alone. Alignments with high *p* values and low ranks should be considered of questionable reliability. Table 1 reveals that sequences of five kilobases or longer in length were usually aligned using the highest ranking alignment. Failure to do so may indicate regions of the *E. coli* W3110 chromosome that are inaccurately mapped or that differ substantially from other strains of *E. coli*.

All of the genes aligned to the physical map in this study already had known genetic map locations. We can use the information obtained from these alignments to assess the utility of computer generated alignments for gene mapping. The five smallest probes aligned were all less than 1400 bp in length but had 4 to 8 sites. All but one of these five had significance values of above 97% and all have probe densities (sites per kilobase of DNA) higher than the average of 2.6 restriction sites per 1 kb (1224 sites/466.2 kb). Since we can cut DNA sequence with any number of enzymes to create probes, the genomic mapping of additional sites (using more than eight enzymes) would have led to an increased probe site density. This would have allowed us to align more DNA sequences. However, inherent redundancy in the chromosome physical maps might present a limitation to this approach. Application of this method to much larger genomes without genetic map information might be impractical since it should take much longer DNA sequences to make the correspondingly more information-rich probes that would be required.

An integrated genomic map of *E. coli* would be useful for refining the genetic map and placing newly sequenced or mapped genes. Unmapped genes can be quickly and analytically aligned using either sequence or restriction map data. Using published data for several unmapped genes, we have been able to locate their most likely map positions . A computerized integrated genomic map can be easily updated to incorporate revised or additional map information. The computerized map enables one to easily keep track of map discrepancies as they accumulate, assisting one in the decision as to whether they are mistakes or reflect strain differences. An integrated map can serve as a framework for organizing other information on the cellular products of *E. coli*, perhaps including such information as two-dimensional gel electrophoresis coordinates, metabolic pathway , and enzymatic properties . The computerized map can be used to model the chromosomes of different *E. coli* strains. For example, we have already simulated the inversion of a segment of the chromosome to eliminate the bothersome IN(*rrnD-rrnE*)*1* inversion from the computerized map (see above). Using the GenBank entries LAMBDACG and ECOBIO (17′), we have also located the bacteriophage lambda attachment site at 819.3 kb and incorporated a sequence-derived lambda restriction map to produce a representation of a lysogenic chromosome. The software tools we developed to analyze the *E. coli* genome should be useful to researchers engaged in the study of the genomes of other organisms. The current embryonic form of the integrated *E. coli* genomic map is a set of files containing the digital genomic restriction map and a growing number of aligned probes. Software has been developed (C. Werner and K. Rudd, personal communication) that produces maps displaying all or any portion of the genomic restriction map with positions of the aligned probes accurately displayed in a variety of scales and formats. We are currently creating a relational database containing integrated *E. coli* map, sequence, reading frame, clone and reference information. Used in conjunction with our map-making software, this database will realize our vision of an integrated *E. coli* genomic map. The software used in this study was written in the C computer language. The data files and analytical software are available on request.

ACKNOWLEDGEMENTS

This project was supported by significant contributions from a number of colleagues. D. Lipman has contributed to all aspects of this project since its inception. We would like to thank and acknowledge the other members of the National Center for Biotechnology Information who have written software for this project: R. Huntzinger, D. Johnson, C. Tolstoshev, and S. Satterfield. We are very grateful to G. Church for providing and updating an annotated compilation of *E. coli* DNA sequence database entries. C. Werner has enthusiastically modified his PostScript based Plasmid Description Language (PDL) to allow us to produce a variety of graphic map representations. Finally, we acknowledge encouraging and enlightening conversations with S. Altschul, M. Berlyn, J.-P. Bouche, G. Church, S. Gottesman, C. Hill, Y. Kohara, G. Olsen, J. Roth, T. Schneider, M. Waterman, and H. Westerhoff. None of this would have been possible without the many mappers and sequencers of the *E. coli* genome who supplied all of the data we analyzed.

REFERENCES

1. Kohara, Y., Akiyama, K., and Isono, K. (1987) Cell, **50**, 495–508.
2. Bachmann, B.J. (1983) Microbiol. Rev., **47**, 180–230.
3. Kroger, M. (1989) Nucleic Acids Res., **17** (supplement), r283–r309.
4. Church, G.M. and Kieffer-Higgins, S. (1988) Science, **240**, 185–188.
5. Pearson, W.R. and Lipman, D.J. (1988) Proc. Natl. Acad. Sci. U.S.A., **85**, 2444–2448.
6. Williams, K.M. (1988) CABIOS, **4**, 211.
7. Waterman, M.S., Smith, T.F. and Katcher, H.L. (1984) Nucleic Acids Res., **12**, 237–242.
8. Zehetner, G. and Lehrach, H. (1986) Nucleic Acids Res., **14**, 335–349.
9. Neumaier, P.S. (1986) Nucleic Acids Res., **14**, 351–362.
10. Sulston, J., Mallett, F., Staden, R., Durbin, R., Horsnell, T. and Coulson, A. (1988) CABIOS, **4**, 125–132.
11. Gumbel, E.J. (1962) In Sarhan, A.E. and Greenburg, B.G. (ed.), Contributions to Order Statistics. Wiley, New York, pp. 56–93.
12. Karlin, S., Ost, F. and Blaisdell, B.E. (1989) *In* Waterman, M.S. (ed.), Mathematical Methods for DNA Sequences. CRC Press, Boca Raton, pp. 133–157.
13. Arratia, R., Gordon, L. and Waterman, M.S. (1986) Ann. Stat., **14**, 971–993.
14. Gordon, L., Schilling, M.F. and Waterman, M.S. (1986) Prob. Th. Rel., **72**, 279–287.
15. Altschul, S.F. and Erickson, B.W. (1986) Bull. Math. Biol., **48**, 617–632.
16. Becker, R.A. and Chambers, J.M. (1984) S: An Interactive Environment for Data Analysis and Graphics. Wadsworth and Brooks/Cole, Pacific Grove.
17. Bachmann, B.J., Low, K.B. and Taylor, A.L. (1976) Bacteriol. Rev., **40**, 116–167.
18. Bachmann, B.J. and Low, K.B. (1980) Bacteriol. Rev., **44**, 1–56.
19. Sanderson, K.E. and Roth, J.R. (1988) Microbiol. Rev., 52, 485–532.
20. Hill, C.W. and Harnish, B.W. (1981) Proc. Natl. Acad. Sci. U.S.A., **78**, 7069–7072.
21. Tabata, S., Higashitani, A., Takanami, M., Akiyama, K., Kohara, Y., Nishimura, Y., Nishimura, A., Yasuda, S. and Hirota, Y. (1989) J. Bacteriol., **171**, 1214–1218.
22. Lin, H.-C., Lei, S.-P. and Wilcox, G. (1985) Gene, **34**, 111–122.
23. Lin, H.-C., Lei, S.-P. and Wilcox, G. (1985) Gene, **34**, 123–128.
24. Lin, H.-C., Lei, S.-P. and Wilcox, G. (1985) Gene, **34**, 129–134.
25. Lee, N., Gielow, W., Martin, R., Hamilton, E. and Fowler, A. (1986) Gene, **47**, 231–244.
26. Muramatsu, S., Kato, M., Kohara, Y. and Mizuno, T. (1988) Mol. Gen. Genet., **214**, 433–438.
27. Ellwood, M. and Nomura, M. (1982) J. Bacteriol., **149**, 458–468.
28. Lin, R.-J., Capage, M. and Hill, C.W. (1984) J. Mol. Biol., **177**, 1–18.
29. Sadosky, A.B., Davidson, A., Lin, R.-J. and Hill, C.W. (1989) J. Bacteriol., **171**, 636–642.