3450–3455 Nucleic Acids Research, 1994, Vol. 22, No. 17

ECD — a totally integrated database of *Escherichia coli* K12

Ralf Wahl, Peter Rice[1], Catherine M.Rice[1] and Manfred Kröger*
Institut für Mikrobiologie und Molekularbiologie, Fachbereich Biologie, Justus-Liebig-Universität Gießen, Frankfurter Straße 107, D-35392 Gießen and [1]EMBL, Postfach 10.2209, Meyerhofstr.1, D-69117 Heidelberg, Germany

ABSTRACT

We have compiled the DNA sequence data for *E.coli* available from the GENBANK and EMBL data libraries and independently from the literature. Starting with this update of our *Escherichia coli* database (ECD release 20) we provide major changes compared to previous issues. This update not only represents another substantial increase in sequence information, it also allows now to find the exact physical location of each individual gene or regulatory region, even regarding descrepancies in nomenclature. In order to save space this printed version does not contain the database itself anymore, but we provide several examples. The complete database is publically available in electronic form together with a self explaining application program or as a flat file. The complete compilation including a full set of genetic map data and the *E.coli* protein index can be obtained in machine readable form from the EMBL data library as a part of the CD-ROM issue of the EMBL sequence database, released and updated every three months. After deletion of all detected overlaps a total of 2 878 364 individual bp is found to be determined till the end of June 1994. This corresponds to a total of 60.98% of the entire *E.coli* chromosome consisting of about 4,720 kbp. This number may actually be higher by 9161 bp derived from other strains of *E.coli*.

INTRODUCTION

Within this sequence supplement issue we were able to publish a compilation of DNA sequences of *Escherichia coli* in five contiguous years since 1989 and asked our colleagues from all over the world for additions and corrections [1–5]. Over the recent years the number of newly published *E.coli* sequence data increased substantially (see Fig. 1). The velocity of adding new data increased also (see Table 1). A rough calculation allows the prediction, that the complete sequence of *E.coli* may be known by 1998, using a noncoordinate effort only. This target may be reached earlier, because at least two groups have devoted their research to systematic sequencing of certain areas of the *E.coli* chromosome. According to our data a total of 2 878 364 bp is sequenced till June 1994. Almost one half of these nucleotides is published more than once. Our database may serve as a basis for encouragement to our colleagues to either send us their unpublished, mostly flanking material or to determine additionally the sometimes very small gaps towards known neighbouring sequences to finally get the complete sequence.

A major aim of our largely increased and totally integrated ECD database is the attempt to provide the entire knowledge about the model organism *Escherichia coli* K12 in an electronic form. We try to use the DNA sequence as the stream line for all other information. Since there are already a number of special databases on different aspect of the *E.coli* cell, we prefer to provide a platform for these different data, only, rather than to build up entirely new system. We allow an unchanged incorporation of data from other databases and prefer to act as a distributer, only.

This compilation is available in its full form quarterly from EMBL data library in electronic form [6]. It may also be received from the EMBL data library on CD-ROM together with an application program for quick database search and direct access to collected sequences.

PREVIOUS AND SUPPORTING EFFORTS

The most famous collection of *E.coli* data is the linkage map compiled by B.Bachmann [7]. These data were updated for the last time in 1990 and we tried to follow this update as close as possible. Thus the electronic full version of ECD contains all known genes, including open reading frames with no genetic function assigned to up till now. Our own literature survey allows us to carry on with the Bachmann collection, at least in part.

Three other groups [8–15] started a program to fit the DNA sequence data directly into the physical map as compiled by Y.Kohara *et al.* [16]. We prefer genetic map positions rather than the physical map coordinates, since this seems to be the easier way to find the most important genetic cross references within the Bachmann map. Nevertheless, we also examine the bigger contiguous sequences (contigs) for fitting into the physical map. In general one may obtain the physical data simply by multiplying the genetic 'map' data using a factor of 47.2. This operation needs to pay attention to a large inversion within the Kohara restriction map, which however is considered in two other cosmidbanks [17, 18]. In order to merge our data with those regarding the Kohara map data directly [8–12] we have compared our data with those of K.Rudd [9] and paid attention

*To whom correspondence should be addressed

to all other collections as far as possible. Thus, we also included the unpublished material available exclusively to us or to K.Rudd [9]. However, these data were given to us for statistical purposes only; for an example of our statistical analysis see references 19 and 20. The respective information may be available on request from the authors themselves.

Although the sequence is not finished now, a number of refined functional analyses of the sequence data are performed, namely on promoter [21] and terminator structures [22], as well as on ribosomal binding sites [23] and on the distribution of REP sequences [24]. Other most valuable collections are available, which regard tRNAs [25], distribution of insertion elements [26, 27], chromatographic behavior of proteins [28] or metabolic pathways [29–30].

SYSTEMATIC SEQUENCING EFFORTS

1992 the first two reports on systematic sequencing of the *E.coli* chromosome were released [31, 32]. According to announcements made on international conferences, these projects have reached a fairly advanced status, already. One group headed by F.Blattner (Madison, WI) has almost finished the systematic sequencing of about 23 % of the genome located between min 77 and 100 [33, 34]. This group has performed their sequencing

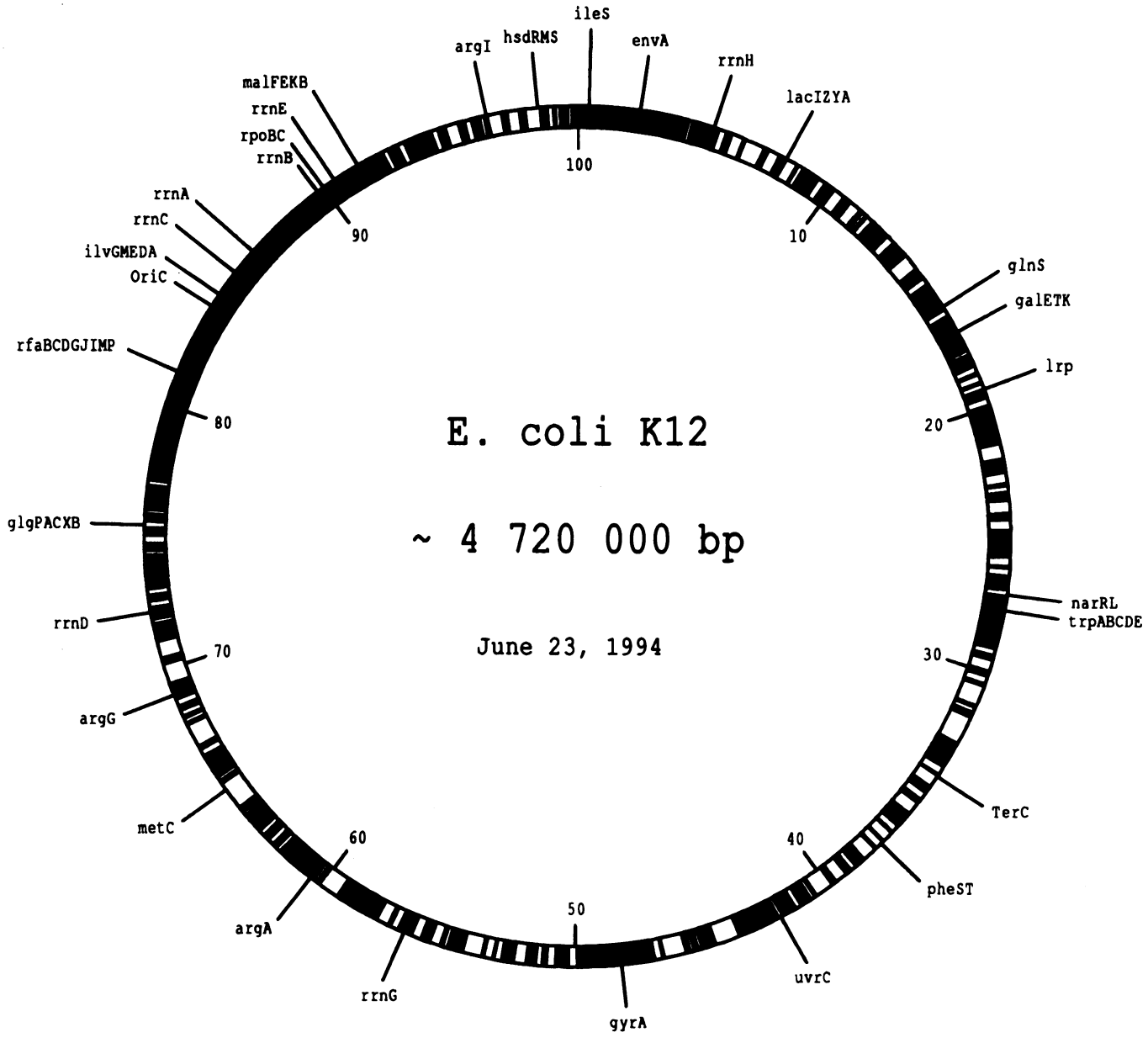

Figure 1. To date sequenced areas within the circular chromosome of *Escherichia coli* K12. The sequenced areas are calculated in percent of the total chromosome. All contigs over 2000 bp are shown in scale black bars at the respective genetic map position. Some prominent markers are included as well. The areas between min 0 and 33, and between min 76 and 100 are subject to systematic sequencing efforts in Japan and the U.S.A., respectively (see text).

independently from any sequence data known from other laboratories. Thus, their first report was upon 91.4 kb with about 75% sequenced in other laboratories before. The published reports already span a segment of 725 112 bp and is beside the yeast chromosomes the largest contiguous sequence reported up till now [35].

Another group is formed by a number of independent laboratories in Japan. They are devoted to systematic sequencing of the area between min 0 and 33. The coordinator is M.Mizobuchi (Tokyo). The first two reports are released, a third one is announced for the near future [36, 37]. The Japanese group uses a completely different strategy by sequencing gaps only. According to personal communication, the finished contig contains 274 kb spanning the entire area from min 0 to 6.4. Their first report on 111 kb included about 65 % sequence data from other laboratories [36].

A third group headed by G.Church (Cambridge, MA) has reported two big contiguous sequences as part of their program to establish new methods in automatic DNA sequence determination [38, 39].

PERFORMED COMPILATION

The general scope of this collection is to allow a compilation of all uncoordinated sequence information to finally end up with a complete *Escherichia coli* nucleotide sequence data base, including all sequenced mutants. In order to give a visual impression about the availability of sequence information of *E.coli* DNA we include an appropriate figure. The extent of the black bars in Figure 1 represents the mainly sequenced areas. All sequences with more then 2000 contiguous basepairs are shown.

We introduced B.Bachmann's genetic map data completely and used them to locate both sequenced and unsequenced areas roughly by a tenth of a minute. Fine assortment was by a hundredth of a minute, if the sequences overlap. A hundredth of a minute corresponds to 472 bp, which seems to be a sufficient resolution. If the sequences were mapped in either of the compilations using the Kohara map [8–14], we preferred to use their assignment including the respective orientation. Contigs are only accepted, if either sequence extends over the respective restriction site. This procedure revealed a fairly good correspondance between genetic and physical map data. However, in the area between min 40 and 80 greater diffences up till 3 min are found. Data given in table 1 are used to recalculate the genetic map position for genes not sequenced yet. Numbers higher then 100 refer to DNA sequences, which could not be localized within the chromosome until now. They are either to small to locate them on the physical map with the necessary confidance or the physical map coordinates have to be changed at the respective position.

```
ID   ompA, tolG, tut, con
AC   EG1102;
XX
DE   E. coli ompA gene (codes for the outer membrane protein II*).
XX
KW   phage recognition
KW   membrane protein
XX
OS   Escherichia coli K12
XX
RL   Nucleic Acids Res. 8:3011 (1980)
RL   Proc. Natl. Acad. Sci. U.S.A. 77:3845 (1980)
RL   J. Biol. Chem. 255:27 (1980)
RL   J. Mol. Biol. 143:317 (1980)
RL   FEBS Letters 128:186 (1981)
RL   Proc. Natl. Acad. Sci. U.S.A. 80:358 (1983)
RL   Genes Dev. 6:135  (1992)
RL   J. Mol.Biol. 176:431 (1984)
RL   Mol. Gen. Genet. 188:472 (1987)
XX
XX
FT   FUNCTION      structural gene
FT   MAP           021.97
FT   CDS           V00307, (1037-2074)
FT   EMBL          V00307, ECOMPA (2271)
FT                 X66232, ECOMPA5PA (136)
FT   CONTIG        ECD021.95, (1037-2074)
FT   SW            P02934
FT   2D            F028.0
XX
SQ   Sequence, 1041 nc, 347 cd, 37201 D,
     atgaaaaaga cagctatcgc gattgcagtg gcactggctg gtttcgctac cgtagcgcag
     gccgctccga aagataacac ctggtacact ggtgctaaac tgggctggtc ccagtaccat
     gatactggtt tcatcaacaa caatggcccg acccatgaaa accaactggg cgctggtgct
     tttggtggtt accaggttaa cccgtatgtt ggctttgaaa tgggttacga ctggttaggt
     cgtatgccgt acaaaggcag cgttgaaaac ggtgcataca aagctcaggg cgttcaactg
     accgctaaac tgggttaccc aatcactgac gacctggaca tctacactcg tctgggtggc
     atggtatggc gtgcagacac taaatccaac gtttatggta aaaaccacga caccggcgtt
     tctccggtct tcgctggcgg tgttgagtac gcgatcactc ctgaaatcgc tacccgtctg
     gaataccagt ggacgaacaa catcggtgac gcacacacca tcggcactcg tccggacaac
     ggcatgctga gcctgggtgt ttcctaccgt ttcggtcagg gcgaggcagc tccagtagtt
     gctccggctc cagctccggc accggaagta cagaccaagc acttcactct gaagtctgac
     gttctgttca acttcaacaa agcaaccctg aaaccggaag gtcaggctgc tctggatcag
     ctgtacagcc agctgagcaa cttggatccg aaagacggtt ccgtagttgt tctgggttac
     accgaccgca tcggttctga cgcttacaac cagggtctgt ccgagcgccg tgctcagtct
     gttgttgatt acctgatctc caaaggtatc ccggcagaca agatctccgc acgtggtatg
     ggcgaatcca acccggttac tggcaacacc tgtgacaacg tgaaacagcg tgctgcactg
     atcgactgcc tggctccgga tcgtcgcgta gagatcgaag ttaaaggtat caaagacgtt
     gtaactcagc cgcaggctta a
//
```

Figure 2. Example ASCII file representing one single ECD entry (for details see text). Each gene or functional element is assigned to such an ASCII file.

```
ID   ECD004.82      standard; DNA; PRO;   17326 BP.
XX
OS   Escherichia coli
XX
AC   D12649; start:     7920 >; length:   4481 nc; (ECD12649)
AC   D12650; start:    12394 >; length:   3929 nc; (ECD12650)
AC   D15061; start:     1786 >; length:   6134 nc; (ECRNHK12)
AC   J01858; start:    12481 >; length:     38 nc; (EC10S3)
AC   K00766; start:    12011 >; length:    247 nc; (ECRGNDS3)
AC   M32357; start:        1 <; length:   2791 nc; (ECDRPA)
AC   M97858; start:        1 <; length:   2795 nc; (ECPROS)
AC   X60739; start:    14739 <; length:   2596 nc; (ECDNIR)
AC   X67217; start:     6807 >; length:     92 nc; (ECRRNHUP)
XX
DE   M97858;  CC: 1;  From   2795;  To   1011;  Offset:       1.
DE   D15061;  CC: 2;  From      1;  To   6134;  Offset:    1786.
DE   D12649;  CC: 3;  From      1;  To   4481;  Offset:    7920.
DE   D12650;  CC: 4;  From      7;  To   3919;  Offset:   12401.
DE   X60739;  CC: 5;  From   1012;  To      1;  Offset:   16314.
XX
SQ   Sequence:   17326 BP.
     aagcttatca ctatgcttca tcgcttttttt aagggaaacg atctggacgc cagcgcctgt
     cgcttctgca acgttattat cacccagaaa gattgcaacg tgactcacag aggaagtgct
     gaagacgcgg attccaaatg aggttacccc aaggcttgag gagaacagca aatcgccggg
     tttgagatca ggtgccgtta tttcttttat tgattgttcg gtgaaagagc tttgatgctg
     gaatttgaca gcccatgttt ttgcctcggc atctactgct gtggctgatg ggtctggctg
     gctgatatca acggtacaag ccgaaagtaa aagaaagctg gggagaaaca ggcggcagta
     cgcctttggt ttatccattt tatacaatcc atgtaaaaaa agggccctga aattcaggac

......

     tttgttccac tgctgcaaat ctttggtgct tacgccgaga cgtgaagcga tacttgaaag
     cgtgtgcgca gagcgtacgg tgtaaacacg gctgttaagc ggcgtattgt cggcaaccag
     cgtcgactgt acagcagcaa tttcgcctga agccagagat tcacgcagtt gatctgcatg
     cttctttggc accatcacgt actgcgggcc acttgcgccc agcgtggagc cttttacgcc
     agcgttgaat gtcttcagct tgctgacgga aatccccgcc atatctgcta ccttcgccat
     ttcaaccggg ctgctcaggt gcacacgcgc cagagcacgg ctttcatcgg tcgttggcag
     acgtacgcca taacgcttgc tgtttttgag aatatcactc aatgccagca ttttaggcac
     gtactgcttc gtttctgcgg caacggtaac gaccagaatg ccgtggattt cccacgcgct
     ttgttcgttt taattgcctt catgacccga ccttcgccgc tgttataagc cgctacggtc
     agaagccagt cgccgtcaaa cattttgttc agacgctgca tcatattcag cgcggcagtt
     gttgaagcaa caacatcgcg acgcgcgtca taattgcgcg tctgtttcaa accataattg
     cgccccgtgc tcggaatgat ctgccagatg cctgcggcat tggcgccaga cgttgcgtga
     ggatcaaaag cgctctccac tatgggtagt agtaccagtt ccataggcat gttacgtttt
     ttaacttgcc ctgctatcca gtacatatac ggctctgccc gtaaagttac atcgtggaga
     tagctcttat tcgtaaatat ttctgtttct gttcgcgaat ccggtcattt tccgaattcc
     catctttaag ctcgtcgcca atgaaagccc acaagtcacc atctgcgcga tagacgtccc
     atcgtccatc catcgtgcct gacttgtaaa ctttgctgct tcccc
//
```

Figure 3. Example for a sequence contig collected from the most actual EMBL database according to the ECD assignments. The DE line is the source of the assignment, while the AC lines indicates all EMBL files matching this contig. The start number represents the position of the respective first nucleotide within the collected contig. The length as well as the genetic orientation ('>' normal orientation; '<' inverse orientation) are indicated and used for the electronic build up of the genetic map as given in Fig. 4. Note, the sequence is shown in part, only.

The gene symbols are preferentially according to the Bachmann list or are taken from a recent publication. However, since there are a number of biases and an increasing number of alternative gene symbols, we have changed our administration program accordingly. Each gene can be found under its historic or systematic name, as well as under the rational name. Each

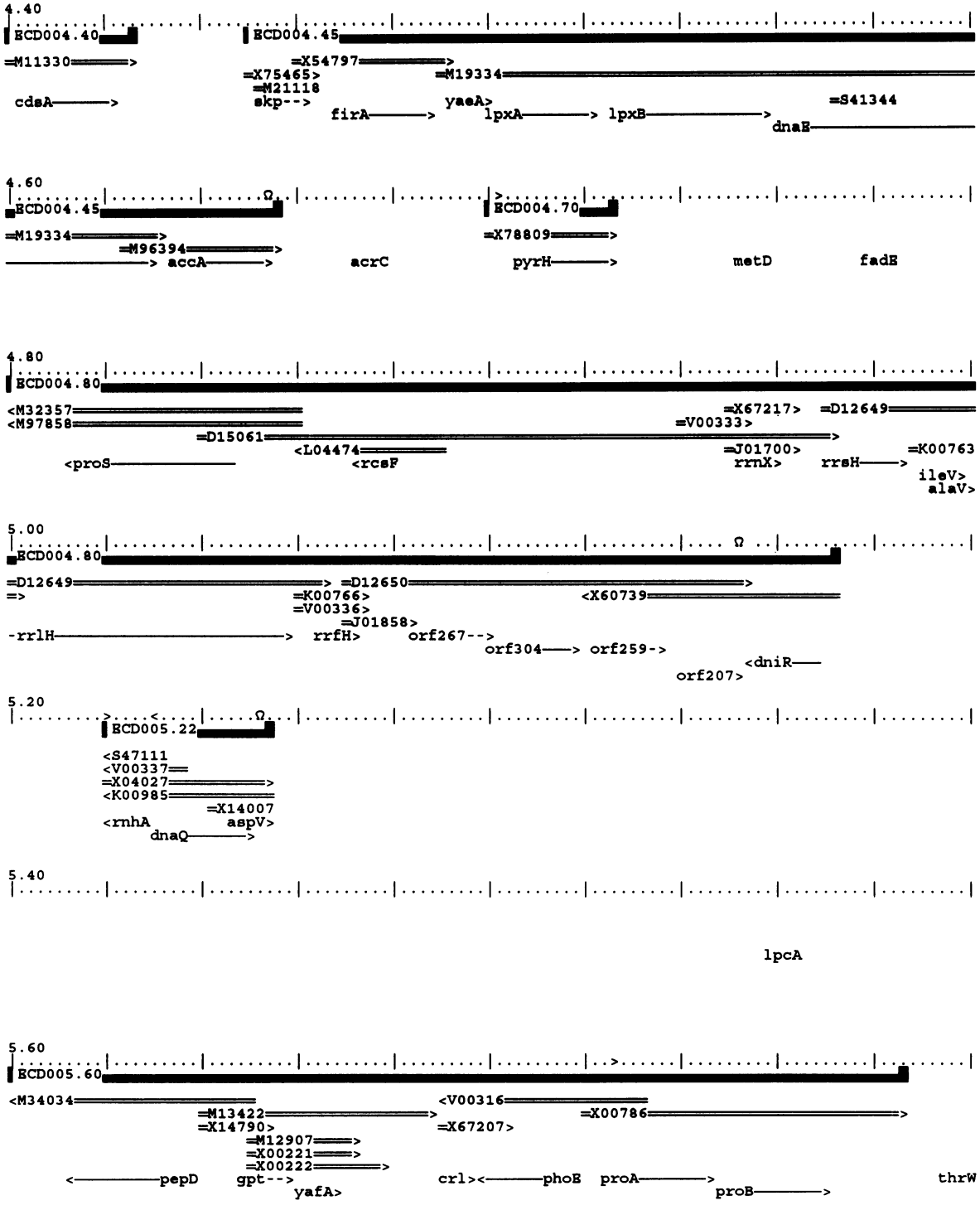

Figure 4. Print out of *E.coli* genetic and physical map in ASCII format. The example given covers the region between min 4.40 and min 5.60. Black bars represent contiguous DNA sequences from more than one file (contigs ECD004.40, ECD004.80, ECD005.60) or from individual EMBL entries (ECD004.40, ECD004.70, ECD005.22). Respective EMBL entry names are given below the black bars, indicating orientation and extent. All bars are drawn to scale. Names of fully assigned genes are accompanied by small arrows indicating the direction of transcription. Genes not sequenced until now are located according to the information given in the Bachmann linkage map. Functional sequences are assigned by 'Ω' for terminators and depending on the orientation either '<' or '>' for promoters within the upper most line.

unannotated open reading frame is named according to the respective publication, but also according to the system propagated by K.Rudd, if already present in EcoSeq6 [9]. Thus although the given entry name may differ sometimes from the EMBL or GenBank entries, an automatic retrival is possible with our new ECD system. Figure 2 gives an example for an ECD entry with four alternative names. A search using the third name 'tut' will automatically lead to the ompA sequence.

Figure 2 may be understood as an example for the new file architecture. In principle we use the same structure as the EMBL data library. Each gene can be retrieved by such an individual file and possesses an individual ECD accession number. Thus our database can now be used directly for cross references using just this number, e.g. in the WWW science net.

This type of file is not only provided for structural genes (ECD system number EGxxxx) but also for specific functional sites (EFxxxx), promoter (EPxxxx), terminator or hairpin structures (EHxxxx), tRNAs (ETxxxx), ribosomal RNAs (ERxxxx) or unannoted open reading frames (EOxxxx). The last type of system numbers are supposed to be completed by an EGxxxx number gradually, as soon as open reading frames are assigned to a known function. Together with a short description line and a line on metabolic function (if known), the keywords derived from different databases are included. A list of cross references is read out in the style of EMBL data library. The feature table (FT) contains all information collected from various databases as well as the calculated map position. Thus references to the Neidthard 2D-protein gel index, to the list of EC-numbers or metabolic pathway index, or to the Brookhaven data base may be found in the features section. The given nucleotide sequence is the most actual sequence excluding any regulatory or flanking sequences. The feature table gives a detailed description of the source of this sequence. Corrections introduced - if necessary - are decribed individually.

This sixth edition of ECD provides another major advantage in connecting all *E.coli* EMBL entries to contigs of maximum extent, which can be broken down into individual files for proteins, insertion elements, katalytic or transfer RNAs. A search for promoter, terminator or other regulatory structures is possible, as long as these features are found in the respective data files. Future issues of ECD may contain additional information manually added by us.

Each contig is compared with the PRO and UNC files of the EMBL database in order to look for yet undetected overlapping sequences. We are able to calculate the exact position of each individual EMBL file within our contigs. This allows a highly detailed map of multiple sequence entries. Figure 3 gives an example for such a contig, which is derived from nine EMBL files and contains 17.2% sequences determined twice or more. Data collection is, however, from five EMBL files, only. Thus ECD is a very convenient source for statistical analysis of sequencing errors [40].

The full set of information is provided in electronic form, which also includes some structural information and other functional data, restriction map data, corrections or sequenced mutations. In addition to the files given in Figures 2 and 3, we are able to provide a genetic map both in electronic form as part of the application program, as well as in printed form. An example for such a print is given in Figure 4. Special symbols are used to illustrate the orientation of individual genes and the presence of promoter and terminator sequences. This print out may be available on special request.

DATA DISTRIBUTION IN MACHINE READABLE FORM

This compilation is available as a set of flat files from the EMBL data library [6] and is automatically distributed with each release the EMBL data library. In addition, this compilation is available on the CD-ROM version of the EMBL data library. This CD-ROM is produced in cooperation with IRL-Press and contains the other collections of this supplement issue, too. However the actual Version or specific sets of data is also available on disk or CD-ROM on request from Gießen, directly. Email address KROEGER@EMBL-HEIDELBERG.DE. or RALF.WAHL@-MIKRO.BIO.UNI-GIESSEN.DE

ACKNOWLEDGEMENTS

We like to thank Kenn Rudd (Bethesda) for his unpublished listing, and Peter Stoehr, David Hazeldine and Rainer Fuchs (EMBL) for constant flow of recent database additions. This work is supported by the Deutsche Forschungsgemeinschaft (Kr 591/7-1).

REFERENCES

1. M.Kröger (1989) Nucl.Acids Res. **17** (Suppl.), r283–309.
2. M.Kröger, R.Wahl and P.Rice (1990) Nucl.Acids Res. **18** (Suppl.) 2549–2587.
3. M.Kröger, R.Wahl and P.Rice (1991) Nucl.Acids Res. **19** (Suppl.) 2023–2043.
4. M.Kröger, R.Wahl, G.Schachtel and P.Rice (1992) Nucl. Acids Res. **20** (Suppl.) 2119–2144.
5. M.Kröger, R.Wahl and P.Rice (1993) Nucl.Acids Res. **21** (Suppl.) 2973–3000.
6. C.M.Rice, R.Fuchs, D..G.Higgins, P.J.Stoehr and G.N.Cameron (1993) Nucl.Acids Res. **21** (Suppl.), 2967–2971.
7. B.J.Bachmann (1990) Microbiol. Rev. **54**, 130–197.
8. K.E.Rudd, W.Miller, J.Ostell and D.A.Benson (1990) Nucl. Acids Res. **18**, 313–321.
9. K.E.Rudd (1992) *In* 'A Short Course in Bacterial Genetics: A Laboratory Manual and Handbook for *Escherichia coli* and Related Bacteria', J.H.Miller (ed.), pp. 2.3–2.43 Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
10. C.Médigue, J.P.Bouché, A.Hénaut and A.Danchin (1990) Molecular Microbiology **4**, 169–187.
11. C.Médigue, A.Hénaut and A.Danchin (1990) Molecular Microbiology **4**, 1443–1454.
12. C.Médigue, A.Viari, A.Hénaut and A.Danchin (1991) Molecular Microbiology **5**, 2629–2640.
13. H.Watanabe and T.Kunisawa (1990) Protein Seq. & Data Analysis **3**, 149–156.
14. T.Kunisawa, M.Nakamura, H.Watanabe, J.Otsuka, A.Tsugita, L.-S.L.Yeh, D.G.George and W.C.Barker (1990) Protein Seq. & Data Analysis **3**, 157–162.
15. C.Médigue, A.Viari, A.Hénault, and A.Danchin (1993) Microbiol. Reviews **57**, 623–654.
16. Y.Kohara, K.Akiyama and K.Isono (1987) Cell **50**, 495–508.
17. R.P.Birkenbihl and W.Vielmetter (1989) Nucleic Acids Res. **17**, 5057–5069.
18. V.Knott, D.J.Blake and G.G.Browlee (1989) Nucl.Acids Res. **17**, 5901–5912.
19. M.McClelland, A.S.Bhagwat, H.-J.Fritz, R.Merkl and M.Kröger, (1992) Nature **355**, 595–596.
20. R.Merkl, M.Kröger, P.Rice and H.-J.Fritz (1992) Nucleic Acids Res. **20**, 1657–1662.
21. S.Lisser and H.Margalit (1993) Nucleic Acids Res. **21**, 1507–1516
22. Y.d'Aubenton Carata, E.Brody and C.Thermes (1992) J.Mol.Biol. **216**, 835–858.
23. K.E.Rudd and T.Schneider (1992) *In* 'A Short Course in Bacterial Genetics: A Laboratory Manual and Handbook for *Escherichia coli* and Related Bacteria', J.H.Miller (ed.), pp. 17.19–17.45 Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1992.
24. G.P.Dimri, K.E.Rudd, M.K.Morgan, H.Bayet and G.Ferro-Luzzi Ames (1992) J.Bacteriology **174**, 4583–4593.

25. S.Steinberg, A.Misch and M.Sprinzl (1993) Nucl.Acids Res. **21** (Suppl.) 3011–3015.
26. M.Umeda and E.Othsubo (1989) J.Mol.Biol. **208**, 601–614.
27. R.P.Birkenbihl and W.Vielmetter (1989) Mol.Gen.Genet. **220**, 147–153.

Table 1. Concordance of physical and genetic map data

Bachmann map	Physical map	Bachmann map	Physical map
0.0 min	0.0 min	55.0 min	57.75 min
5.0 min	4.90 min	60.0 min	62.90 min
10.0 min	10.00 min	65.0 min	67.90 min
15.0 min	14.55 min	70.0 min	72.70 min
20.0 min	20.55 min	75.0 min	76.35 min
25.0 min	25.33 min	80.0 min	80.00 min
30.0 min	30.30 min	85.0 min	85.30 min
35.0 min	35.65 min	90.0 min	90.10 min
40.0 min	40.90 min	95.0 min	94.90 min
45.0 min	46.20 min	100.0 min	100.00 min
50.0 min	52.35 min		

28. R.A.VanBogelen, P.Sanker, R.L.Clark, J.A.Boga and F.C.Neidhardt (1992) Electrophoresis **13**, 1014–1054.
29. M.Riley (1993) Microbiol. Reviews **57**, 862–952.
30. E.Selkov (1994) personal communication - information available via SELKOV@SHOME.ITEB.SERPUKHOV.SU
31. D.L.Daniels, G.Plunkett III, V.Burland and F.Blattner (1992) Science **257**, 771–778.
32. V.Burland, G.Plunkett III, D.L.Daniels, and F.Blattner (1993) Genomics, **16**, 551–561.
33. G.Plunkett III, V.Burland, D.L.Daniels, and F.Blattner (1993) Nucleic Acids Res. **21**, 3391–3398.
34. F.Blattner, V.Burland, G.Plunkett III, H.J.Sofia, and D.L.Daniels (1993) Nucleic Acids Res. **21**, 5408–5417.
35. H.J.Sofia, V.Burland, D.L.Daniels, G.Plunkett III, and F.Blattner (1994) Nucleic Acids Res. **22**, 2576–2586.
36. T.Yura, H.Mori, H.Nagai, T.Nagata, A.Ishihama, N.Fujita, K.Isono, K.Mizobuchi and A.Nakata (1992) Nucleic Acids Res. **20**, 3305–3308.
37. N.Fujita, H.Mori, T.Yura and A.Ishihama (1994) Nucleic Acids Res. **22**, 1637–1639.
38. P.Richterich, N.Laskey, G.Gryan, L.Jaehn, L.Mintz, K.Robison, G.M.Church (1993) EMBL/GenBank AccNo. U00007
39. P.Richterich, N.Laskey, G.Gryan, L.Jaehn, L.Mintz, K.Robison, G.M.Church (1993) EMBL/GenBank AccNo. U00008
40. G.Schachtel, R.Wahl, and M.Kröger, to be published elsewhere.

Table 2. Annual growing of the *E.coli* DNA-sequence information. Part A table compiles all published sequence information according to the annual growth in number of publications (entries) and nucleotides given therein. Part B compiles the respective mumbers and percentage after subtraction of overlaps as annually published in this supplement [1–5].

Part A

Year	annual entries	total entries	annual information (bp)	total information (bp)
1967	1	1	600	600
1968	2	3	205	805
1969	5	8	349	1154
1970	3	11	249	1403
1971	10	21	841	2244
1972	5	26	426	2670
1973	7	33	447	3117
1974	4	37	205	3322
1975	9	46	820	4142
1976	6	52	432	4574
1977	9	61	1116	5690
1978	22	83	6911	12601
1979	44	127	19936	32537
1980	54	181	35375	67912
1981	73	254	90801	158713
1982	99	353	83318	242031
1983	141	494	130996	373027
1984	152	646	178749	551776
1985	179	825	164574	716350
1986	198	1023	248965	965315
1987	204	1227	203736	1169051
1988	221	1448	268866	1437917
1989	255	1703	294480	1732397
1990	637	2340	328475	2060872
1991	328	2668	377251	2438123
1992	481	3149	582913	3021036
1993	539	3688	1120281	4141317
1994	321	4009	453644	4594961

Part B

date	actual information	percentage
January 1989	940 449 bp	19.92 %
February 1990	1 248 696 bp	26.46 %
February 1991	1 492 282 bp	31.62 %
February 1992	1 820 237 bp	38.56 %
April 1993	2 353 635 bp	49.87 %
June 1994	2 878 364 bp	60.98 %