46–49 Nucleic Acids Research, 1998, Vol. 26, No. 1

Compilation of DNA sequences of *Escherichia coli* K12: description of the interactive databases ECD and ECDC (version 1997)

Manfred Kröger* and Ralf Wahl

Institut für Mikrobiologie und Molekularbiologie, Fachbereich Biologie, Justus-Liebig-Universität Gießen, Frankfurter Straße 107, D-35392 Gießen, Germany

Received October 3, 1997; Accepted October 6, 1997

ABSTRACT

We have compiled the DNA sequence data for *Escherichia coli* K12 available from the GenBank and EMBL data libraries and independently from the literature. We provide the most definitive version of the ECD *Escherichia coli* database now exclusively via the World Wide Web System (http://susi.bio.uni-giessen.de/ecdc.html). Our database encloses the completed genome sequence recently published by two competing groups and an assembled set of all elder sequences. The organisation of the database allows precise physical location of each individual gene or regulatory region, even taking into consideration discrepancies in nomenclature. The WWW program allows to the user to branch into the original EMBL and SWISS-PROT datafiles. A number of links to other WWW servers dealing with *E.coli* is provided. A FASTA and BLAST search may be performed online. Besides the WWW format a flat file version may be obtained via ftp. A number of discrepancies between the two systematic sequence determinations and/or the literature have not yet been resolved. However, our database may serve as a reference source for resolution and/or the assignment of strain difference.

INTRODUCTION

Within this database issue we have been able to publish a compilation of DNA sequences of *Escherichia coli* for eight consecutive years since 1989 and colleagues from all over the world have provided additions and corrections (1–8). The primary target, namely to submit the complete sequence of *E.coli* K12 to either of the international databanks was reached independently by two groups on December 28, 1996 and on January 16, 1997. We began to spread this message on January 23, 1997 (9,10). Meanwhile, the sequence data were released completely, and published after some updating (11,12). After completion of the DNA sequence the full functional description of this major model organism is the next target, which is a long way from completion. Thus, this compilation needs to be kept for sometime, and it has gained rather than lost it's importance.

AIM OF COMPILATION

The aim of our *E.coli* database (ECD) is to provide an electronic entry into the entire knowledge about the model organism *E.coli* K12. We use the DNA sequence as the basis for all other information. Since there are already a number of specialist databases on different aspect of the *E.coli* cell, we prefer to provide a platform for these different data, rather than to build up an entirely new system. We allow unchanged incorporation of data from other databases and prefer to act only as a distributor. In order to make this point as clear as possible, our World Wide Web system is called ECDC for *E.coli* database collection (13).

For previous and supporting efforts please see our previous papers in this series and the references quoted therein (6,8,13). Mainly because the international databases provide powerful computer programs for homology searches across their entire set of data, an increasing number of people will pick up homologies to *E.coli* genes. Consequently, there is a great demand for a reliable and accurate database of *E.coli* genomic data, especially for data on hypothetical open reading frames (ORFs). However, collecting all the acquired data in one database is very difficult for individual laboratories. Thus ECDC and the ECD nucleotide collection therein tries to provide a service for all other databases dealing with *Escherichia coli*. The World Wide Web system seems to be an ideal tool to connect different databases, which are maintained directly at the original laboratories. Any comments and corrections are very welcome at the addresses given below.

Although we have often been asked to also collect data from pathogenic strains, we have restricted ourself to *E.coli* K12. Instead, we provide a number of links to other WWW servers, which may lead into other databases collecting these data.

PERFORMED COMPILATION

After the completion of the entire sequence no means are needed to bridge any gaps within the recently updated genetic map (14). Only a few genes described have not yet been assigned to defined map positions. Theses can be found in ECD in a separate line (see Fig. 1).

Since a number of smaller contigs could not be localized within the chromosome of the two sequenced strains, and since they are not described as epichromosomal genes, we refer to them using

*To whom correspondence should be addressed. Tel: +49 641 99 35530; Fax: +49 641 99 35549; Email: kroeger@bio.uni-giessen.de

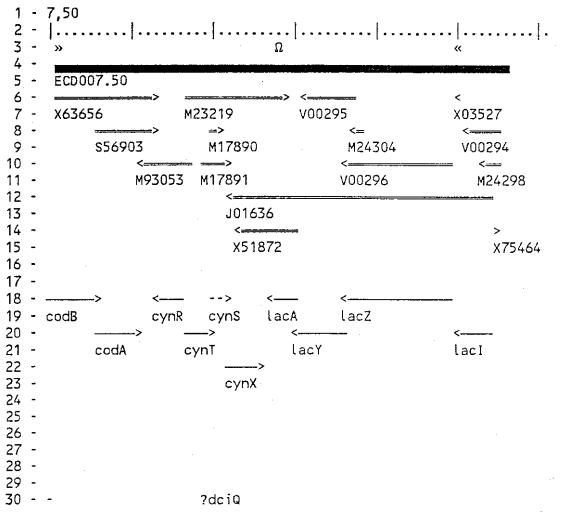

Figure 1. Organization of the ECDC interactively usable genetic map. All bars are drawn to scale. Names of fully assigned genes are accompanied by small arrows indicating the direction of transcription. Genes not sequenced until now are located according to the information given in the Bachmann linkage map. Functional sequences are assigned by 'Ω' for terminators and depending on the orientation either '<' or '>' for promoters within the uppermost line. For a more detailed description see ref. 13.

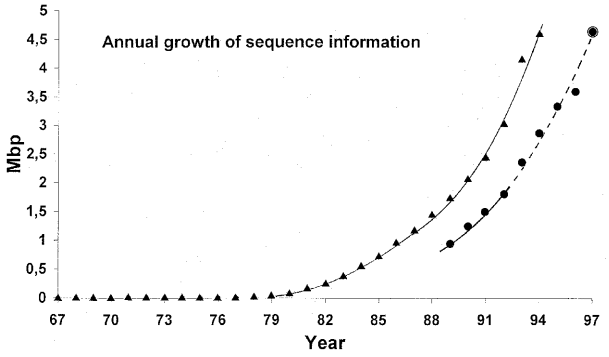

Figure 2. Annual growth of sequence information. Data originally given in refs 6, 7 and 8 are plotted either as uncorrected (triangles) or as corrected (dots) sequence information. All points form a line, which in 1993 (28) was used as a basis to predict, that the end point (circled dot) may be reached in 1997. The broken line represents the prediction. As illustrated, the actual data followed this prediction very closely.

map positions >100. We provide 77 unmapped sequences with a total of 140 966 bp. It still seems to be very important to collect these entries, since during the course of systematic sequencing it became very clear that we were dealing with a number of different K12 substrains. A very good example is the new outer membrane-associated protease OmpP, which could not be mapped on any of the Kohara phages (15). Using artificial map positions allows to include all additional sequences of this type into our FASTA or BLAST searches.

The gene symbols are preferentially according to the recent genetic map (14) or are taken from a most recent publication. However, since there are a number of biases and an increasing number of alternative gene symbols, we have made up our administration program accordingly. As long as no international agreement is found, each gene can be found under its historic or systematic name, as well as under the rational name. Thus, although the given entry name may sometimes differ from the EMBL or GenBank entries, an automatic retrival for alternative names is possible with our ECD system. For an example see figure 2 of our previous update (6). Searches may also be performed using keywords in the near future.

The hypothetical ORFs are especially hard to deal with. Each unannotated ORF is named according to the respective publication, but also listed according to the system propagated by K.Rudd, if already present in EcoSeq7 (14) or clearly annotated in the respective EMBL/GenBank entry. A search mode using the recently introduced *b*-numbers of *b*lattner's group (11) will be introduced in either of the next releases. We have not used the system introduced by the group from Japan (12). The system is used for two thirds of the genome only, and it refers to the artificial phage numbering system. People interested in obtaining certain phages are asked to look into the respective area of the www-page of the Japanese group [http://genome4.aist-nara.ac.jp/].

A number of ORFs are not found in the systematic sequence. Since they are part of both literature (16–18) and recent genetic map (14), we have incorporated them into ECD. In most cases they are annoted in SWISS-PROT, but not in the respective nucleotide database files. Thus it is often not clear which nucleotide changes have to be introduced. Areas in question are easily detectable, since in almost all of these cases more than one ORF is shown for the same area. Since we tried to provide all of these sometimes very doubtful predictions, and since the most recent updates are not yet incorporated, our numbers of genes and ORFs slightly deviate from those given by the systematic sequencing projects. We keep 4215 protein-coding genes, while Blattner *et al*. report 4288 (11). We keep 1916 characterized genes, while GenBank keeps 1827 of them. Blattner *et al*. do not provide such a figure. Anyhow, we have to expect that these numbers will change for a while. Since there is a constant flow of newly described functions, please refer always to the most recent release of ECDC. We try to monitor the current literature as quickly as possible.

Figure 2 of our previous update (6) may also be understood as an example for the respective file architecture. In principle, we use the same structure as the EMBL data library. Each gene can be retrieved as an individual file and possesses an individual ECD accession number. Thus our database can be used directly for cross references using just this number, e.g., in the World Wide Web system.

Individual files are 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 ORFs (EOxxxx). The last type of system numbers are supposed to be replaced by an EGxxxx number gradually, as soon as ORFs 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 the 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 2D-protein gel index (19), to the list of EC-numbers or metabolic pathway index (20–22), to the New

Haven *E.coli* Genetic Stock Center (23) or to the Brookhaven database may be found in the features section. The respective links are provided with hypertext and allow a direct entry into the respective databases. 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. Compiling the sequence data in this way allows monitoring of every correction performed by the respective author automatically.

ECD provides the *E.coli* sequence in 20 individual contigs. This size has been used for a number of releases and people seem to be happy with it. Both the entire sequence and the 400 individual Blattner files, as well as the sequence broken down into individual phage sequences as provided by the Japanese group may be obtained by clicking onto the appropriate EMBL/GenBank AccNr. All ECD contigs provide individual files for proteins, insertion elements, catalytic or transfer RNAs. Thus both handling and homology search are quick and easy. A search for promoter, terminator or other regulatory structures is possible, as long as these features are described in the respective data files. Future issues of ECD may contain additional information, e.g., keywords, manually added by us. Please refer always to the most recent release.

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 individual data files, we provide a genetic map in electronic form as part of the application program (13). An example for the interactively usable genetic is given in Figure 1. Special symbols are used to illustrate the orientation of individual genes and the presence of promoter and terminator sequences. Gene symbols and EMBL accession numbers are provided as hypertext. Gene symbols will provide the individual feature table together with the nucleotide sequence of the gene. Hypertext within the feature table of the ECD entry as well as within the EMBL file allows a most convenient connection to other databases, e.g., to MEDLINE abstracts.

SOME RESULTS OF STATISTICAL ANALYSIS

During the course of the data collection, it became evident that many features of the entire *E.coli* sequence may be predictable with high confidence. We like to give the most prominent examples, which could be confirmed by systematic sequencing.

A chain of ATGA-interlinked genes turned out to be the most surprising feature of the completed sequence of bacteriophage lambda (24). The same feature has been described independently in our laboratory with a fragment of the phage sequence (25). Ever since we have been interested in the function of this curious overlap. Consequently, we have inspected the *E.coli* data since 1982. This marked the start of the ECD collection presented here. Each case found resulted in a special note to our public ECD files. To our satisfaction, Blattner *et al*. report about the ATGA overlap as one of the most interesting features of the entire sequence (11).

Using ~20% of the sequence information we pointed out that the CTAG tetramer is extremely rare in *E.coli*. We were able to first present a feasible explanation for this on the *E.coli* conference held at the Banbury Center of Cold Spring Harbor Laboratories, NY, in October 1991 (26,27). Blattner *et al*. stress the importance of this result (11).

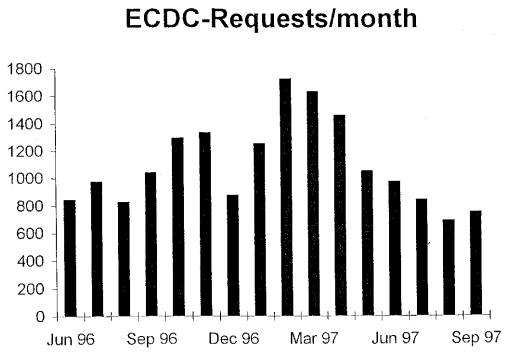

Figure 3. Frequency of visitors of ECDC homepage. Individual visitors are counted on a daily, and summarized on a monthly, basis. The highest frequency was reached immediately after the release of the entire sequence, which could be introduced into the ECDC system in March 1997.

During the second international *E.coli* conference (Madison, 1993) we presented a sketch plotting the increase of sequencing information (see table 1 in ref. 6). The plot given in Figure 2 was used to predict that the sequence may be finished by 1997 (28). It is hard to believe that the sequence not only is finished once at the predicted time, but that it is independently covered 2.96 times.

Most areas of the *E.coli* genome are sequenced three times, some even four times, and even five times is not exactly rare. We used this rich source of data for an analysis of sequencing error distribution. Although these statistics have to be repeated with the entire data set, we are able to present three preliminary rules (29). (i) The most likely sequencing error is the omission or addition of an extra nucleotide in a row of four or five or even more identical nucleotides (homopolynucleotide rule). (ii) GC inversions are very likely to occur (GC-rule). (iii) In elder sequences the flanking areas contain significantly more errors than internal sequences. There are two sharp borders <30 and <150 nucleotides, which may be due to cloning artefacts and to sequencing into only one direction, respectively.

DATA DISTRIBUTION IN MACHINE READABLE FORM AND INTERNATIONAL ACCEPTANCE

Since we discontinued the distribution on CD-ROM or on paper, the only way to use the ECD *Escherichia coli* database is via the ECDC database collection on the World Wide Web (WWW) system. Besides a simple and fairly approximate statistical analysis of user identification (see below) we do not read or collect any data submitted for a search within ECD. Users of our ECD database or our ECDC database collection should use the URL: http://susi.bio.uni-giessen.de/ecdc.html . They are politely asked to cite this paper within scientific publications and/or grant applications.

Computer programs freely available within the WWW system allow a fairly detailed statistical analysis about the user identification and frequency. Since June 1996 more than 17 600 individual researchers used our system. This corresponds to an average of 37 different individuals using ECDC each day. According to Figure 3, soon after the completion of the entire sequence, interest in ECDC was highest, gradually going back to a normal frequency over the last few months. Users of ECDC reside in more than 45

different countries. For more information, use either of these email addresses kroeger@bio.uni-giessen.de or wahl@fmi.ch.

ACKNOWLEDGEMENT

We would like to thank the staff at EBI (Cambridge) for the constant flow of recent database additions.

REFERENCES

1 Kröger, M. (1989) *Nucleic Acids Res.*, **17** (Suppl.), r283–309.
2 Kröger, M., Wahl, R. and Rice, P. (1990) *Nucleic Acids Res.*, **18**, 2549–2587.
3 Kröger, M., Wahl, R. and Rice, P. (1991) *Nucleic Acids Res.*, **19**, 2023–2043.
4 Kröger, M., Wahl, R., Schachtel, G. and Rice, P. (1992) *Nucleic Acids Res.*, **20**, 2119–2144.
5 Kröger, M., Wahl, R. and Rice, P. (1993) *Nucleic Acids Res.*, **21**, 2973–3000.
6 Wahl, R., Rice, P., Rice, C.M. and Kröger, M. (1994) *Nucleic Acids Res.*, **22**, 3450–3455.
7 Kröger,M. and Wahl,R. (1996) *Nucleic Acids Res.*, **24**, 29–31.
8 Kröger, M. and Wahl, R. (1997) *Nucleic Acids Res.*, **25**, 39–41.
9 Hobom, B.(1997) *Frankfurter Allgemeine Zeitung*, Nat.&Wiss., February 5, 1997.
10 OBrien, C.(1997) *Nature*, **385**, 472.
11 Blattner, F.R., Plunkett, G.,III, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, Ch.K., Mayhew, G.F. *et al.* (1997) *Science*, **277**, 1453–1462.
12 Yamamoto, Y., Aiba, H., Baba, T., Hayashi, K., Inada, T., Isono, K., Itoh, T., Kimura, S., Kitagawa, M., Makino, K. *et al.* (1997) *DNA Res.*, **4**, 91–113.
13 Wahl, R. and Kröger, M. (1995) *Microbiol. Res* (Jena), **150**, 7–61.
14 Berlyn, M.B., Low, K.B. and Rudd, K.E. (1996) In Neidhardt,F.C. (ed.) *Escherichia coli and Salmonella–Cellular and Molecular Biology*, 2nd edition. pp. 1715–1902, ASM Press, Washington, DC.
15 Kaufmann, A., Stierhof ,Y.-D. and Henning, U. (1994) *J. Bacteriol.*, **176**, 359–367.
16 Koonin, E.V., Tatusov, R.L. and Rudd, K.E. (1996) In Neidhardt,F.C. (ed.) *Escherichia coli and Salmonella–Cellular and Molecular Biology*, 2nd edition. pp. 2203–2217, ASM Press, Washington, DC.
17 Borodovsky, M., Koonin, E.V. and Rudd, K.E. (1994) *Nucleic Acids Res.*, **22**, 4756–4767.
18 Borodovsky, M., Koonin, E.V. and Rudd, K.E. (1994) *Trends Biochem. Sci.*, **19**, 309–313.
19 VanBogelen, R.A., Abshire, K.Z., Pertsemlidis, A., Clark, R.L. and Neidhardt ,F.C. (1996) In Neidhardt,F.C. (ed.) *Escherichia coli and Salmonella–Cellular and Molecular Biology*, 2nd edition. pp. 2067–2117, ASM Press, Washington, DC.
20 Karp, P.D., Riley, M., Paley, S.M., Pelligrini-Toole, A. and Krummenacker, M.(1997) *Nucleic Acids Res.*, **25**, 43–50. [See also this issue, *Nucleic Acids Res.* (1998), **26**, 50–53.]
21 Riley, M. (1997) *Nucleic Acids Res.*, **25**, 51–52. [See also this issue, *Nucleic Acids Res.* (1998) **26**, 54.]
22 Riley, M. and Labedan, B. (1996) In Neidhardt,F.C. (ed.) *Escherichia coli and Salmonella–Cellular and Molecular Biology*, 2nd edition. pp. 2118–2202, ASM Press, Washington, DC.
23 Berlyn, M.B. and Letovsky, S. (1992) *Nucleic Acids Res.*, **20**, 6143–6151.
24 Sanger, F., Coulsen, A.R., Hong, G.F., Hill, D.F. and Hill, G.B. (1982) *J. Mol. Biol.*, **162**, 729–773.
25 Kröger, M. and Hobom, G. (1982) *Gene*, **20**, 25–38.
26 McClelland, M., Bhagwat, A.S., Fritz, H.-J., Merkl, R. and Kröger, M. (1992) *Nature*, **355**, 595–596.
27 Merkl, R., Kröger, M., Rice, P. and Fritz, H.-J. (1992) *Nucleic Acids Res.*, **20**, 1657–1662.
28 Kröger, M. and Wahl, R. (1993) International E.coli Genome Meeting, Madison, Wisconsin, September 9–13, 1993. Abstract book, p. T7, 2.
29 Schachtel, G., Wahl, R. and Kröger, M., unpublished results.