

PromEC: An updated database of *Escherichia coli* mRNA promoters with experimentally identified transcriptional start sites

Ruti Hershberg, Gill Bejerano[1], Alberto Santos-Zavaleta[2] and Hanah Margalit*

Department of Molecular Genetics and Biotechnology and [1]School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel and [2]Centro de Investigación sobre Fijación de Nitrógeno, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México

Received September 13, 2000; Accepted October 3, 2000

ABSTRACT

PromEC is an updated compilation of *Escherichia coli* mRNA promoter sequences. It includes documentation on the location of experimentally identified mRNA transcriptional start sites on the *E.coli* chromosome, as well as the actual sequences in the promoter region. The database was updated as of July 2000 and includes 472 entries. PromEC is accessible at http://bioinfo.md.huji.ac.il/marg/promec

INTRODUCTION

We announce the availability of an updated compilation of *Escherichia coli* mRNA promoter sequences (as of July 2000). The database extends our previous compilation (1), and as in the original compilation, contains only promoter sequences with experimentally identified transcriptional start sites. The original database has been used extensively both by researchers who are studying the properties of promoter sequences and their relationship to gene expression, and by computational biologists who have used it as a benchmark for testing newly developed prediction algorithms on a reliable data set. We believe that the updated database will be as valuable to both communities.

DATABASE CONTENT AND ORGANIZATION

The added promoter sequences are based on three sources: the documentation in the *E.coli* genome file (2), the RegulonDB database (3) and an updated literature search (covering the years 1993–2000). Only promoters with documented transcriptional start sites were extracted from the genome file and the RegulonDB database. Promoters that were compiled following the literature search were those for which the transcriptional start sites were identified experimentally, either by primer extension or by S1 nuclease mapping. RegulonDB database has also incorporated these updates based on the literature search. The promoters in our original compilation were also re-examined in view of the documentation in the *E.coli* genome file and the RegulonDB database. As a result, a few sequences, deemed as incorrect, were removed or corrected. In most cases, those involved second or third mRNA start sites of the same gene.

The updated database currently includes 472 entries: 275 sequences from the original compilation, 157 promoters collected from the other two databases and 40 additional promoters that were found in publications dated between the years 1993 and 2000. The database documents the gene name, the position of the transcriptional start site on the *E.coli* genome, the reference paper and the sequence itself (spanning nucleotides –75 to +25 relative to the transcriptional start site). If more than one start site was documented for a single gene, the most prominent one is recorded in the position field (if such information is available), while all start sites are marked in the sequence record.

AVAILABILITY OF THE DATABASE

The database is available at http://bioinfo.md.huji.ac.il/marg/promec, where a flat file containing all promoter sequences may also be extracted. When citing the database please cite this reference, along with the RegulonDB database (3). We expect the database to grow as new transcriptional start sites are determined. It is also probable that we have missed some already known transcriptional start sites. Any information of that kind would be very much appreciated and should be sent to promec@md2.huji.ac.il

ACKNOWLEDGEMENTS

We thank Julio Collado-Vides and Heladia Salgado for their help. This study was supported by grants from the Human Frontiers Science Program (H.M.), and from the Ministry of Science, Israel (G.B.).

REFERENCES

1. Lisser,S. and Margalit,H. (1993) Compilation of *Escherichia coli* mRNA promoter sequences. *Nucleic Acids Res.*, **21**, 1507–1516.
2. Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
3. Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millan-Zarate,D., Blattner,F.R. and Collado-Vides,J. (2000) RegulonDB (version 3.0): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 65–67. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 72–74.

*To whom correspondence should be addressed at: Department of Molecular Genetics and Biotechnology, The Hebrew University—Hadassah Medical School, POB 12272 Jerusalem 91120, Israel. Tel.: +972 2 6758614; Fax: +972 2 6784010; Email: hanah@md2.huji.ac.il