

Genes and proteins of *Escherichia coli* (GenProtEc)

Monica Riley* and David B. Space

Marine Biology Laboratory, Woods Hole, MA 02540, USA

Received August 18, 1995; Revised and Accepted October 16, 1995

ABSTRACT

GenProtEc is a database of *Escherichia coli* genes and their gene products, classified by type of function and physiological role and with citations to the literature for each. Also present are data on sequence similarities among *E.coli* proteins with PAM values, percent identity of amino acids, length of alignment and percent aligned. The database is available as a PKZip file by ftp from mbl.edu/pub/ecoli.exe. The program runs under MS-DOS on IBM-compatible machines. GenProtEc can also be accessed through the World Wide Web atURL http://mbl.edu/html/ecoli.html.

GenProtEc (Genes and Proteins of *Escherichia coli*) is a dBase style database and associated query program for IBM-compatible PCs which centers around the gene products of *E.coli*. As of July 1993 the database contained 1717 gene products whose physiological function was known to some degree (1), some better understood than others. As of Fall 1994 the number had risen to 1894 (2). Additions continue to be made.

There are two main sets of data, one for genes, gene products and their physiological role, the other identifying *E.coli* proteins of similar sequence. The gene/gene product database contains the full gene product name, the Enzyme Commision (EC) number for enzymes, the gene name and synonyms, the type of gene product and the category of physiological function of the gene product. Up to three literature references are supplied for each entry.

All gene products have been classified as to type, as either an enzyme, a regulator, RNA, part of the membrane, a member of a transport system, a protein factor, a carrier or a part of the structure of the cell other than membrane. The distribution of 1894 *E.coli* gene products by type has been summarized (3). The gene products have also been assigned to one or up to four of 118 categories of physiological function. An early version of the classification of genes and gene products by function is in Riley (1), a more recent version will be available soon (3).

One can search this database by gene name or a synonym of the gene name, by gene product string or by physiological category. Complete pick lists are available for each of these. The search can be refined by adding more terms with the logical relationships AND, OR and AND/OR.

Information on sequence similarities within the *E.coli* genome is also available. When a sequence similarity exists between the amino acid sequence of any chosen gene product and the sequence of another *E.coli* protein information about the similarity is available. The database contains the results of similarity analyses (2) that used AllAllDB of the Darwin suite at Zurich (4), requiring an alignment of at least 100 amino acids and a PAM (accepted point mutations) score (5) of <250. The 1894 *E.coli* K-12 chromosomally encoded proteins of known sequence formed 2126 pairs with sequence similarity as defined above. The major characteristics of the 2126 pairs reside in GenProtEc. Again, pick lists are available for the SWISS-PROT mnemonic names. After selecting the starting sequence the names of other *E.coli* proteins that have a sequence similar to it are shown. Selecting any one, the length of the alignment of the two proteins is shown together with the percent of the protein aligned, the percent identical amino acids and the PAM score. As new sequences are provided by the SWISS-PROT database (6) additional information on seqence relationships will be incorporated into the database.

The database and associated query program are available at no charge as a self-extracting compressed binary file on the anonymous site pub/ecoli/ecoli.exe through the Marine Biological Laboratory's server hoh.mbl.edu. The user name is 'anonymous' and the password is the user's email address. Once downloaded, GenProtEc may be expanded by the command 'ecoli'. Following expansion, the query program may be run by entering the command 'eco'. The database can be queried directly on the World Wide Web, accessing through the home page at URL http://www.mbl.edu/html/Riley/Monica.html.

Feedback will be gratefully received. Users kindly cite this article.

REFERENCES

1 Riley,M. (1993) *Microbiol. Rev.*, **57**, 862–952.
2 Labedan,B. and Riley,M. (1995) *Mol. Biol. Evol.*, **12**, 980–987.
3 Riley,M. and Labedan,B. (1996) In Curtiss,R., Lin,E.C.C., Ingraham,J., Low,K.B., Magasanik,B., Neidhardt,F., Reznikoff,W., Riley,M., Schaechter,M. and Umbarger,H.E. (eds), *Escherichia coli and Salmonella.* American Society for Microbiology, Washington, DC, in press.
4 Gonnet,G.H., Cohen,M.A. and Benner,S.A. (1992) *Science*, **256**, 1443–1445.
5 Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) In Dayhoff,M.O.(ed.), *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Washington, DC, Vol. 5, suppl. 3, pp. 345–358.
6 Bairoch,A. and Boeckman,B. (1993) *Nucleic Acids Res.*, **21**, 3093–3096.

* To whom correspondence should be addressed