

Genes and Proteins of *Escherichia coli* K-12 (GenProtEC)

Monica Riley*

Marine Biology Laboratory, Woods Hole, MA 02540, USA

Received October 15, 1997; Accepted October 24, 1997

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, representing groups of paralogous genes, with PAM values, percent identity of amino acids, length of alignment and percent aligned. GenProtEC can be accessed at the URL http://www.mbl.edu/html/ecoli.html

GenProtEC (Genes and Proteins of *E.coli*) is a database available on the World Wide Web which centers around the products of *Escherichia coli* K-12 chromosomal genes. The database contains a listing of 4403 genes, 2086 gene products whose physiological function is known to some degree empirically, some better understood than others, 1244 open reading frams (ORFs) whose function can be predicted by sequence similarity with known proteins and 1073 ORFs of unknown function. The data base contains the gene name, its synonyms, the SwissProt (1) mnemonic for proteins when one has been assigned, its synonyms, the full gene product name, and the Enzyme Commision EC number for enzymatic reactions. Up to three literature references are supplied for each entry.

Data on physiological function and sequence similarity are also given. Gene products have been classified as to type, as either an enzyme, a regulator, RNA, part of the membrane, a member of the transport system, a protein factor, a carrier, or a part of the structure of the cell other than membrane. The gene products have been assigned to at least one or up to four of 118 hierarchically arranged categories of physiological function (2,3).

Sequence similarity of each protein to any other *E.coli* protein is given, permitting the grouping together of *E.coli* proteins of similar amino acid sequence. The database contains the results of similarity analyses carried out in collaboration with Bernard Labedan (4,5). using the AllAllDB of the Darwin suite at Zurich (6), requiring an alignment of at least 100 amino acids and a PAM score (accepted point mutations) (7) of <200. Almost half of *E.coli* K-12 chromosomally encoded proteins had at least one *E.coli* protein partner with sequence similarity as defined above. Some proteins were essentially fusions of two or more independent proteins that we term 'modules'. To avoid artifact and error, proteins consisting of two or more modules >100 amino acids each were divided, so that each module was treated separately. The resulting 2149 proteins/domains formed 7161 sequence-related pairs. The pairs were linked by chains of similarities into sequence-related groups. There are 602 sequence-related groups of *E.coli* proteins, ranging in size from 2 to 129, and most or all members of each group are related by function as well as by sequence.

One can query GenProtEC with a gene name or a synonym or with a SWISS-PROT name or a synonym, or with a string for description of gene product or a key for physiological category. Complete pick lists are available for each of these. Information on the gene product and the function of the gene product is returned, as well as sequence similarities among *E.coli* proteins. For any protein that has at least one sequence-related partner, the name(s) of all other *E.coli* proteins in the related group are returned. For any sequence-related pair, the position and length of the alignment for each of the two proteins is given, as well as the percent of the protein aligned, the percent identical amino acids and the PAM score.

The database can be queried directly on the World Wide Web, accessing through the URL http://www.mbl.edu/html/ecoli.html . Feedback and corrections will be gratefully received. Users are requested to kindly cite this article.

ACKNOWLEDGEMENTS

Grateful thanks to David Space and David Remsen, Information Sevices Division, Marine Biological Laboratory, for invaluable programming and site design.

REFERENCES

1 Bairoch,A. and Boeckman,B. (1993) *Nucleic Acids Res.*, **21**, 3093–3096. [See also this issue *Nucleic Acids Res.* (1998) **26**, 38–42].
2 Riley,M. (1993) *Microbiol. Rev.*, **57**, 862–952.
3 Riley,M. and Labedan,B. (1996) In Curtiss,R.,III, 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, pp. 2118–2202.
4 Labedan,B. and Riley,M. (1995) *Mol. Biol. Evol.*, **12**, 980–987.
5 Riley,M. and Labedan,B. (1997) *J. Mol. Biol.*, **268**, 857–868.
6 Gonnet,G.H., Cohen,M.A. and Benner,S.A. (1992) *Science*, **256**, 1443–1445.
7 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.

*Tel: +1 508 289 7612; Fax: +1 508 540 6902; Email: mriley@mbl.edu