Completing the E. coli proteome: a database of gene products characterised since the completion of the genome sequence

Gavin H. Thomas

Department of Molecular Microbiology, John Innes Centre, Norwich, Norfolk NR4 7UH, UK

Received on April 21, 1999; revised on May 19, 1999; accepted on May 21, 1999

Abstract
Summary: A database collating research on E. coli genes whose products have been characterised subsequent to in silico predictions from the completed genome sequence.
Availability: The database is available on the WWW at http://sun1.bham.ac.uk/bcm4ght65/genome.html.
Contact: gav.thomas@bbsrc.ac.uk

The bacterium *Escherichia coli* has been intensively studied over the last 50 years and is arguably the most well understood of all free-living organisms. The completion of the genome sequence was a landmark event in understanding the biology of this bacterium. However, the information gleaned from the genome sequence illustrated that, despite the reams of papers published concerning *E. coli*, there is still much to learn. Of the 4405 predicted genes, only 2200 have been characterised to any extent, leaving approximately 50% of the potential coding capacity uncharacterised. The task of *E. coli* biologists is to complete this knowledge by determining the functions of these gene products. When this has been achieved we will be closer to completely understanding the biology of this model organism. Within these additional 2200 gene products will be proteins with important functions in higher organisms as well as proteins which may be important in the pathogenicity of other bacteria.

In order to keep up to date with this process, a database has been created to record the functions of new gene products as they are determined. These data are primarily gathered from recent papers in the scientific literature. The database was started in September 1998 after the release of the most recent edition of the *E. coli* linkage map (Berlyn, 1998). Updates are posted regularly on the 'What's new' page of the 'E. coli index', of which this database forms a part (Thomas and Bettelheim, 1998; http://sun1.bham.ac.uk/bcm4ght6/res.html).

Each entry in the database lists the new and old gene names, including the Blattner number, and links to the SWISS-PROT entry for the translation of each gene (Bairoch and Apweiler, 1997). The map position is shown in minutes and kb on the Blattner sequence and the genetic organisation of each gene is also included. The identified function of the gene product is indicated with a key reference for further information. Each newly characterised operon has an additional line containing the operon name, the overall function of the gene products and a reference to a key researcher in this area. Wherever possible, links to home pages and Email addresses of these researchers are provided. A number of example entries are illustrated in Figure 1.

The database will grow rapidly over the coming years and should finally reach around 2200 entries when all the potential open reading frames have been determined. In theory there are as many genes yet to be characterised as have already been characterised and at the current rate of about 10 identifications per month the project should be complete sometime in the next twenty years.

References


<table>
<thead>
<tr>
<th>New gene name</th>
<th>Old gene name</th>
<th>Swiss Prot</th>
<th>Location</th>
<th>Function of product</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>tatB</em> (mitA)</td>
<td>mgT (b3838)</td>
<td>None</td>
<td></td>
<td>Component of translocase</td>
<td><a href="http://WWW">WWW</a>. Joel Weiner Email: <a href="mailto:Joel.Weiner@ualberta.ca">Joel.Weiner@ualberta.ca</a></td>
</tr>
<tr>
<td><em>tatC</em> (mitB)</td>
<td>mgUV (b3839)</td>
<td>P27857</td>
<td></td>
<td>Essential component of translocase</td>
<td>Sargent <em>et al.</em> (1998) <em>EMBO J</em> 17:3640-3650. (Characterisation of tatABCD and tatB)</td>
</tr>
<tr>
<td><em>tatD</em> (mitC)</td>
<td>mgW (b3839)</td>
<td>P27859</td>
<td></td>
<td>Essential component of translocase</td>
<td><a href="http://WWW">WWW</a>. Tracy Palmer Email: <a href="mailto:PALMER@bbsrc.ac.uk">PALMER@bbsrc.ac.uk</a></td>
</tr>
<tr>
<td><em>udhA</em></td>
<td>(b3962)</td>
<td>P27306</td>
<td>Single gene at minute 89.5 (4158.3 kb).</td>
<td>Soluble pyridine nucleotide transhydrogenase</td>
<td>Boons <em>et al.</em> (1999) <em>J Bacteriol</em> 181:1030-1034. <a href="http://WWW">WWW</a>. Neil Bruce Email: <a href="mailto:N.Bruce@biotech.cam.ac.uk">N.Bruce@biotech.cam.ac.uk</a></td>
</tr>
</tbody>
</table>

Fig. 1. A section of the database illustrating an example for (i) an operon (the *tat* operon) and (ii) a single gene (the *udhA* gene).