BIOINFORMATICS

The Mammalian Gene Mutation Database

P.D.Lewis1, J.S.Harvey, E.M.Waters and J.M.Parry

Centre for Molecular Genetics and Toxicology, School of Biological Sciences, University of Wales Swansea, Swansea SA2 8PP, UK

The Mammalian Gene Mutation Database (MGMD) is a comprehensive collection of published mutation data from the open literature on mammalian cell-based gene model mutation detection systems. The database currently contains approximately 30,000 comprehensively described mutant spectra records and it is maintained and updated on a daily basis. The major objectives of the MGMD were (i) to provide an Internet-accessible database (http://lisntweb.swan.ac.uk/cmgt/index.htm) for chemically induced and spontaneous mutation types and spectra in selected genes; (ii) to standardize the reporting of mutations within different genes where ambiguity exists in the literature; and (iii) to provide interactive and user-friendly access to the information. A multi-option search facility has been included that allows the user to search the database for parameters such as mutagen, gene or cell type of interest. The structure of the database permits easy retrieval of specific mutation data for further analysis. Thus, the MGMD should become a useful and necessary reference source and provides an analysis tool for genetic toxicologists.

Introduction

Whilst it has generally been assumed that many cancers are the result of a series of genetic changes produced by exposure to exogenous agents, our ability to establish causal relationships between chemical exposures and the induction of specific tumour types has proved to be elusive. It is now possible to examine such relationships due to the advances in molecular biology which have enabled the analysis of mutations both in specific target genes associated with disease, such as oncogenes, as well as those induced by specific chemical or physical treatments in model systems (Olsen et al., 1996; Waters et al., 1999; Perera and Weinstein, 2000). These advances have lead to a remarkable increase in the amount of DNA sequence information in the literature regarding both mutation types (profiles) and spectra (the frequencies of mutations observed at each nucleotide position within a gene). This, in turn, led to the requirement for computerized databases to keep pace with the information increase.

The Mammalian Gene Mutation Database (MGMD) is a project designed to collate the profiles and spectra of published mutagen-induced and spontaneous gene mutations from model mammalian cell types following analysis of the literature. Chemically induced mutation profiles and spectra can be used to determine the specificity of mutagen interaction within a target nucleotide sequence (Coller and Thilly, 1994). A knowledge of mutagen interaction with specific DNA sequence may provide a powerful tool in determining the causes of mutation in tumours. One of the rationales of the MGMD was to build a collection of chemical mutation spectra from different mammalian cell types to highlight potential differences in metabolism and mechanisms of mutagenesis. It was also realized that there was a need to overcome the ambiguities in the way mutations are reported for different genes and present the data in a uniform way. The MGMD is an Internet-accessible database which aims to be an up-to-date and complete reference source for mutagen-induced and spontaneous mutational spectra which will be of considerable benefit to researchers in genetic toxicology. Furthermore, during development of such an extensive database it became apparent that access would have to be ‘user friendly’. Therefore, a simple to use, yet comprehensive search page has been developed that allows rapid and specific retrieval of mutant data.

Data coverage

The MGMD currently contains entries for nearly 30,000 mutants for the genes listed in Table I. Data have been collected for those genes used extensively in mutagenesis studies in cells derived from a variety of mammalian species. The quantity of data for the supF gene alone is extensive. For example, mutant sequence data for supF is available for human, monkey, mouse (cell lines and transgenic animals) and rat covering over 10 different tissue types and more than 40 different cell lines. In total, there is a diverse range of mutant data for the genes in Table I for over 115 chemical and physical mutagens. In addition, there is a substantial collection of spontaneous mutation data for most of the genes. Data entry into the MGMD is an ongoing process with new mutant sequences added daily, and it is envisaged this task will be made easier by user submission of data via the Internet site.

Database structure

Before inclusion in the database each publication was evaluated to determine a number of key factors including: experimental method, cell type used, modifications of the cell type, including factors such as DNA repair status, chemical exposure regime, etc. Each individual mutant entry is also cross-referenced to ensure that duplication of entry does not occur. A major effort has been made to standardize the way in which the data are displayed for the user. For each entry into the database it has been important to classify each kind of mutation. In the

1To whom correspondence should be addressed. Tel: +44 1792 205200; Fax: +44 1792 205200; Email: balewis@swan.ac.uk
MGMD, mutants are classed as single base substitution, tandem base substitution (two adjacent singles), multiple base substitution (non-adjacent singles), complex (mutations of more than one type), deletion, insertion and rearrangements.

It was also important to clarify and standardize the position of the mutation within a particular gene where non-uniformity exists in the literature. In general, the principles recommended by Antonarakisi and the Nomenclature Working Group (1998) were adopted. For practical reasons, different numbering systems have been adopted to describe mutations in the various genes of interest. In assigning a nucleotide number to the mutated base or bases, the following points were taken into consideration. Firstly, the numbering system used should be clear and unambiguous; secondly, it should conform to the numbering system generally accepted and used by workers in each gene system. Data entered into the MGMD has been edited, and the nucleotide numbering system standardized according to the following criteria: (i) *supF*, data are reported for the non-transcribed strand (or coding strand) as the genomic DNA position; (ii) *APRT*, data are reported for the non-transcribed strand, where the A of the AUG initiation codon = nucleotide position 1 using the nucleotide sequence for the entire gene with continuous numbering of both exons and introns; (iii) *HPRT*, data are reported for the non-transcribed strand, where the A of the AUG initiation codon = nucleotide position 1. Mutations occurring in exons are numbered according to their cDNA position. Mutations occurring in introns are numbered thus: where I3 + 1 = first base of intron3 and I3 − 1 = last base of intron3. Amino acids are numbered where the Met initiation codon = amino acid 1.

All mutant entries are held within one Microsoft Access™ database book file. Each entry is subdivided into columns housing relevant mutation information: an organization that facilitates the multi-option search facility online.

**Table I.** The genes and numbers of induced and spontaneous mutants currently held within the MGMD

<table>
<thead>
<tr>
<th>Gene</th>
<th>No. of mutants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Induced</td>
</tr>
<tr>
<td>supF</td>
<td>13980</td>
</tr>
<tr>
<td><em>HPRT</em></td>
<td>9541</td>
</tr>
<tr>
<td><em>APRT</em></td>
<td>934</td>
</tr>
<tr>
<td><em>gpt</em></td>
<td>397</td>
</tr>
<tr>
<td><em>dhfr</em></td>
<td>79</td>
</tr>
<tr>
<td><em>lucI</em></td>
<td>407</td>
</tr>
<tr>
<td><em>lucZ</em></td>
<td>63</td>
</tr>
<tr>
<td>Total</td>
<td>25401</td>
</tr>
</tbody>
</table>

**Database access**

The online database can be accessed by the user from the URL: http://lisntweb.swan.ac.uk/cmgt/index.htm

The URL leads to a home page allowing the user to select options to search the database, review the database instructions and nomenclature, submit mutant data, access links to other database web sites or view MGMD statistics. The ‘user friendly’ search page contains drop down boxes for each searchable category (Figure 1). An enquiry can be made individually by mutagen, species, cell origin (tissue type), cell type, gene, mutation class (single substitution, deletion, etc.) and mutation type (transition, transversion, etc.). Alternatively, the user can refine their search by selecting options within different categories thus allowing any combination of factors attributable to the mutants to be sought. This system has the advantage of eliminating awkward search terms and the many various ways researchers refer to compounds and cell lines. An additional benefit is an option where the scientist can search by author. The

**Fig. 1.** The MGMD search page with multi-option drop-down boxes. The user can narrow his/her search down by specifying search criteria. It is also possible to search for particular references to display mutant data.
The Mammalian Gene Mutation Database system, therefore, allows the user to obtain a list of only the mutation data that is required using single or multiple choices via the options provided. In addition, the search page provides useful examples of how to obtain specific data.

After submitting a search query the results are displayed in a scrollable webpage of mutation information (Figure 2). The mutant information is presented in columns describing: mutagen, dose, species, cell origin, cell type, gene/vector, mutation type, mutation class, nucleotide number, original sequence, mutated sequence, codon, amino acid change and reference. Each reference is also accompanied by a hyperlink to the corresponding entry in Medline to allow rapid access to the paper’s abstract. The user can then save these pages or highlight the mutation spectra information and cut and paste it into Microsoft Excel™ for further analysis.

Conclusions and future developments

The MGMD is currently the most comprehensive online database containing information on induced mutations in mammalian cells. The need for a continuously updated ‘user-friendly’ interactive site has been recognized which, we believe, will make the MGMD a useful tool for genetic toxicologists. The database will also complement others dedicated to similar and different areas of mutation research such as the Mutation Spectra Database at Yale University (Hutchinson and Donnellan, 1997), the Human Gene Mutation

Fig. 2. These results page displaying the columns describing each mutation entry. These data can be saved or copied and pasted into a spreadsheet for reference or analysis.
Database (Cooper et al., 1998) and the Human p53, HPRT, lacI, lacZ databases developed by Cariello et al. (1998). A future addition to the MGMD will include pages dedicated to generated graphical mutation spectra and profiles from the data which may be downloaded directly by the user.

Acknowledgements
The development of the database has been made possible by the support from the European Union Environmental Programme, Otsuka Pharmaceuticals and BAT Ltd.

References