Databases and software for the analysis of mutations in the human *p53* gene, human *hprt* gene and both the *lacI* and *lacZ* gene in transgenic rodents

Neal F. Cariello*, George R. Douglas[1], Nancy J. Gorelick[2], Daniel W. Hart, John D. Wilson and Thierry Soussi[3]

Glaxo Wellcome, Inc., Medicines Safety Evaluation, 5 Moore Drive, Research Triangle Park, NC 27709, USA, [1]Health Canada, Environmental Health Centre, Mutagenesis Section, Tunney's Pasture, Ottawa, Ontario K1A 0L2, Canada, [2]The Procter & Gamble Co., Ivorydale Technical Center, Cincinnati, OH 45217, USA and [3]UMR 218 du CNRS, Institut Curie, 75248 Paris, France

ABSTRACT

We have created databases and software applications for the analysis of DNA mutations at the human *p53* gene, the human *hprt* gene and both the rodent transgenic *lacI* and *lacZ* loci. The databases themselves are stand-alone dBASE files and the software for analysis of the databases runs on IBM-compatible computers with Microsoft Windows. Each database has a separate software analysis program. The software created for these databases permit the filtering, ordering, report generation and display of information in the database. In addition, a significant number of routines have been developed for the analysis of single base substitutions. One method of obtaining the databases and software is via the World Wide Web. Open the following home page with a Web Browser: http://sunsite.unc.edu/dnam/mainpage.html . Alternatively, the databases and programs are available via public FTP from: anonymous@sunsite.unc.edu. There is no password required to enter the system. The databases and software are found beneath the subdirectory: pub/academic/biology/dna-mutations. Two other programs are available at the site, a program for comparison of mutational spectra and a program for entry of mutational data into a relational database.

INTRODUCTION

We have created databases and software applications for the analysis of DNA mutations at several loci. This brief communication describes databases and software for analysis of mutations at the human *p53* gene, the human *hprt* gene and both the transgenic *lacI* and *lacZ* loci. In addition, the *lacI* database contains information on mutations in the bacterial gene. These loci are of interest for a variety of reasons, as outlined below.

Mutations at one of the loci, the *p53* gene, are found with high frequency in a wide variety of human cancers. The types of mutations that occur in the *p53* gene may provide information

regarding the mechanisms of carcinogenesis and may be related to prognosis. It is estimated that perhaps 50% of all human cancers contain a mutation in the *p53* oncogene (1,2).

A gene of interest to genetic toxicologists is the hypoxanthine guanine phosphoribosyl transferase (*hprt*) gene which codes for an enzyme that functions in the purine salvage pathway. Cells bearing a mutation in the *hprt* gene can be selected and cloned from tissue culture experiments and from T-cells isolated from rodents (3), primates (4) and humans (5,6). Thus somatic mutations arising *in vivo* in humans can be studied and compared.

The development of transgenic rodents for the study of mutation is relatively new. These systems typically employ a transgenic lambda phage shuttle vector and use the *lacI* (7) or *lacZ* (8) genes as mutational targets. These systems permit the analysis of mutations generated *in vivo* in a variety of tissues (9,10).

A considerable amount of information about mutations generated in *Escherichia coli* also exists for the *lacI* and *lacZ* loci. Data about mutations generated in *E.coli* are present in the *lacI* database but not the *lacZ* database. The reason for this is that the primary sequence of the *lacZ* construct used in transgenic animals is not identical to the lacZ sequence in bacteria; compared to the bacterial sequence, the transgenic *lacZ* gene contains a 15 base insert near the 5′ portion of the gene. The software described in this article requires a single DNA sequence as a reference point; multiple sequences cannot be accomodated. Therefore no information about *lacZ* mutations in bacteria is present.

In order to facilitate the analysis of mutations, we have developed databases containing DNA sequence information about the *hprt*, *p53*, *lacI* and *lacZ* genes and a software package that performs summary and statistical analysis of the information in each database. The number of mutations in each database is as follows: p53, 5900; hprt, 2500; lacI, ~1500 transgenic and 8000 bacterial; lacZ, ~400.

DATABASES

Each database is in the dBASE format and is present as a stand-alone file. Information common to all databases includes (i) base position, (ii) the nature of the mutation, (iii) amino acid

position, (iv) wt and mutant amino acid, (v) the local sequence around a mutation and (vi) literature citation.

Information specific to the *p53* database includes (i) cancer type, (ii) cell origin (tumor, cell line, etc.) and (iii) loss of heterozygosity. Data particular to the *hprt* database includes (i) mutagen, (ii) dose, (iii) background and induced mutation frequencies, (iv) information whether the mutant was generated *in vivo* or *in vitro*, (v) mRNA splicing information for mutants affecting splicing and (vi) cell type. Data contained in the *lacZ* transgenic database includes (i) dose, (ii) time from last treatment to animal sacrifice, (iii) supplier, species, strain, sex and age of animal, (iv) the organs selected for mutation analysis, (v) the mutant fraction in each organ, (vi) the total PFU analyzed and (vii) the plaque color. Data in the *lacI* database are very similar to the *lacZ* database.

All databases and software are described in considerable detail in other publications (11–14).

SOFTWARE

A separate software package exists for each database, the software runs on IBM-compatible PCs only. All software packages run under Microsoft Windows 3.1, Windows 95 or Windows NT. All software packages permit the filtering, ordering, report generation and display of information in the database.

A significant number of routines have been developed for the analysis of single base substitutions, including programs to (i) determine if two mutational spectra are different, (ii) display mutable amino acids in the protein, (iii) determine if mutations show a DNA strand bias, (iv) determine the frequency of transitions and transversions, (v) display the number and kind of mutations observed at each base in the coding region and (vi) perform nearest neighbor analysis. For genes with exons, a routine will display the number of mutations and mutable sites in each exon. Graphics displays are available for mutated amino acids and for mutational spectra representation. RasMol, a public domain viewer for protein structure, has been included with the appropriate protein structures for *p53*, *lacI* and *hprt*.

AVAILABILITY

The databases and software for the *p53* and *lacI* and *lacZ* gene are freely available. The *hprt* database and software is available on a subscription basis, however, a version of the database and software is available for evaluation.

One method of obtaining the databases and software is via the World Wide Web (WWW). Open the following home page with a Web Browser: http://sunsite.unc.edu/dnam/mainpage.html

Alternatively, the databases and programs are available via public FTP from: anonymous@sunsite.unc.edu. There is no password required to enter the system. The databases and

software are found beneath the subdirectory: pub/academic/biology/dna-mutations. The FTP server is very popular and users may not be able to get in to the system during peak hours.

Information about all databases and instructions for downloading are present when using either WWW or FTP access. All files must be transferred as binary files.

ADDITIONAL PROGRAMS AVAILABLE

Two other programs are available at the site, a program for comparison of mutational spectra (15) and a program for entry of mutational data into a relational database (16). The mutational spectra program is a stand-alone DOS executable and the relational database program requires Microsoft Access 2.0 to run and modify the program.

The present article is an extension of the work presented in the previous *Nucleic Acids Research* database issues (17–19).

REFERENCES

1 Caron de Fromentel,C. and Soussi,T. (1992) *Genes Chromosomes Cancer*, **4**, 1–15.
2 Hollstein,M., Sidransky,D., Vogelstein,B. and Harris,C.C. (1991) *Science*, **253**, 49–53.
3 Jones,I.M., Burkhart-Schultz,K. and Crippen,T.L. (1987) *Somat. Cell Mol. Genet.*, **13**, 325–333.
4 Harbach,P.R., Filipunas,A.L., Wang,Y. and Aaron,C.S. (1992) *Environ. Mol. Mutagen.*, **20**, 96–105.
5 Albertini,R.J., O'Neill,J.P., Nicklas,J.A., Heintz,N.H. and Kelleher,P.C. (1985) *Nature*, **316**, 369–371.
6 Turner,D.R., Morley,A.A., Haliandros,M., Kutlaca,R. and Sanderson,B.J. (1985) *Nature*, **315**, 343–345.
7 Provost,G.S., Kretz,P.L., Hamner,R.T., Matthews,C.D., Rogers,B.J., Lundberg,K.S., Dycaico,M.J. and Short,J.M. (1993) *Mutat. Res.*, **288**, 133–149.
8 Douglas,G.R., Jiao,J., Gingerich,J.D., Gossen,J.A. and Soper,L.M. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 7485–7489.
9 Gossen,J.A., de Leeuw,W.J., Tan,C.H., Zwarthoff,E.C., Berends,F., Lohman,P.H., Knook,D.L. and Vijg,J. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 7971–7975.
10 Kohler,S.W., Provost,G.S., Kretz,P.L., Dycaico,M.J., Sorge,J.A. and Short,J.M. (1990) *Nucleic Acids Res.*, **18**, 3007–3013.
11 Cariello,N.F. and Gorelick,N.J. (1996) *Environ. Mol. Mutagen.*, **28**, 397–404.
12 Cariello,N.F. and Douglas,G.R. (1996) *Environ. Mol. Mutagen.*, **28**, 145–153.
13 Cariello,N.F., Cui,L., Beroud,C. and Soussi,T. (1994) *Cancer Res.*, **54**, 4454–4460.
14 Cariello,N.F. (1994) *Mutat. Res.*, **312**, 173–185.
15 Cariello,N.F., Piegorsch,W.W., Adams,W.T. and Skopek,T.R. (1994) *Carcinogenesis*, **15**, 2281–2285.
16 Cariello,N. (1996) *Mutat. Res.*, **359**, 103–117.
17 Cariello,N.F., Beroud,C. and Soussi,T. (1994) *Nucleic Acids Res.*, **22**, 3549–3550.
18 Cariello,N.F., Douglas,G.R. and Soussi,T. (1996) *Nucleic Acids Res.*, **24**, 119–120.
19 Cariello,N.F., Douglas,G.R., Dycaico,M.J., Gorelick,N.J., Provost,G.S. and Soussi,T. (1997) *Nucleic Acids Res*, **25**, 136–137.