Nucleic Acids Research, 1998, Vol. 26, No. 1 **205–213**

IARC Database of p53 gene mutations in human tumors and cell lines: updated compilation, revised formats and new visualisation tools

P. Hainaut*, T. Hernandez, A. Robinson[1], P. Rodriguez-Tome[1], T. Flores[1], M. Hollstein[2], C. C. Harris[3] and R. Montesano

International Agency for Research on Cancer, 150 cours Albert-Thomas, 69372 Lyon Cedex 08, France, [1]EMBL-Outstation, European Bioinformatics Institute, Hinxton Hall, Cambridge, UK, [2]German Cancer Research Center, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany and [3]Laboratory of Human Carcinogenesis, National Cancer Institute, Bethesda, MD 20892, USA

Received October 10, 1997; Accepted October 24, 1997

ABSTRACT

Since 1989, about 570 different p53 mutations have been identified in more than 8000 human cancers. A database of these mutations was initiated by M. Hollstein and C. C. Harris in 1990. This database originally consisted of a list of somatic point mutations in the *p*53 gene of human tumors and cell lines, compiled from the published literature and made available in a standard electronic form. The database is maintained at the International Agency for Research on Cancer (IARC) and updated versions are released twice a year (January and July). The current version (July 1997) contains records on 6800 published mutations and will surpass the 8000 mark in the January 1998 release. The database now contains information on somatic and germline mutations in a new format to facilitate data retrieval. In addition, new tools are constructed to improve data analysis, such as a Mutation Viewer Java applet developed at the European Bioinformatics Institute (EBI) to visualise the location and impact of mutations on p53 protein structure. The database is available in different electronic formats at IARC (http://www.iarc.fr/p53/homepage.htm) or from the EBI server (http://www.ebi.ac.uk). The IARC p53 website also provides reports on database analysis and links with other p53 sites as well as with related databases. In this report, we describe the criteria for inclusion of data, the revised format and the new visualisation tools. We also briefly discuss the relevance of *p*53 mutations to clinical and biological questions.

INTRODUCTION

The p53 tumor suppressor gene encodes a nuclear phosphoprotein with cancer-inhibiting properties. The development of human cancer often involves inactivation of this suppressor function through several mechanisms, including loss of alleles at the p53 locus (on 17p13), deletions, insertions, point mutations or silencing of the p53 protein by complex formation with viral or cellular proteins. These mutations frequently arise somatically. However, p53 mutations may also be inherited in some families with a predisposition to multiple cancers, as in the Li–Fraumeni syndrome (LFS). Point mutations are scattered over more than 250 codons and are common in many forms of human cancer. In this respect, the p53 gene differs from other tumor suppressor genes such as Rb, APC, BRCAI and $p16^{MTS-1}$ which are frequently inactivated by deletion or nonsense mutations, and from the oncogenes of the *ras* family, which are activated by mutation at a small number of well-defined codons (reviewed in refs 1–3).

Since the identification of somatic tumor-specific, missense p53 mutations in 1989, there has been a widespread interest in the possibility that the localization and the characteristics of these mutations may reveal clues about the etiology and the molecular pathogenesis of human cancer (4–6). The aim of the IARC p53 mutation database is to provide a tool to classify, sort, retrieve, compare and analyze these mutations in order to generate hypotheses on the natural history of human cancer. Initially started as a simple list of mutations in 1990, the database has been developed as a collective effort by several scientists in Europe and in the United States (7). The database is maintained at the International Agency for Research on Cancer and its structure has been revised to meet the demands of a rapidly growing community of users. Ultimately, the aim of the database is to provide rapid and detailed information on p53 mutations, as well as links with other related databases, in the following areas: (i) molecular epidemiology of cancer, (ii) molecular pathology of human cancer and (iii) structural analysis of p53 protein. In the first part of this report, we provide factual information on the scope of the database, the criteria for inclusion of data, the structure and the formats in which the database is currently available and the tools for data analysis and visualisation. In the

*To whom correspondence should be addressed. Tel: +33 4 7273 8532; Fax: +33 4 7273 8575; Email: hainaut@iarc.fr

second part, we briefly review how analysis of p53 mutations contributes to our understanding of human cancer. The electronic form of the database may be cited by referencing this *Nucleic Acids Research* article.

SCOPE OF THE MUTATION COMPILATION AND CRITERIA FOR INCLUSION OF DATA

The scope of the database is to retrieve and arrange data from the literature in a standardized electronic format. This provides a powerful means for manipulation, comparison, search and retrieval of records describing the nature of *p*53 mutations in various cancers. The database is based exclusively on published p53 mutations associated with human cancers. This includes mutations found in normal, pre-neoplastic and neoplastic tissues, including metastases, as well as in cell lines derived from such tissues. However, the database does not include information on human tissues which are reported as negative with respect to p53 mutation. Thus, the database does not allow the calculation of frequency of occurrence of p53 mutations in human cancers, nor the frequency of specific polymorphisms. Experimentally induced mutations in tumor cells or cell lines *in vitro* and p53 mutations in animal tumors are also beyond the present scope of our task.

In contrast with previous versions of the database, information on mutations identified in the germline as well as on common p53 polymorphisms are now being compiled and will be released in the database updates from January 1998. This information will be made available in the form of two tables (germline mutations, 108 entries and polymorphisms, five entries), distinct from the table of somatic mutations (6598 entries). Information on germline mutations is compiled from a database of 108 different families with germline p53 mutation, initially developed at IARC by P. Kleihues and H. Oghaki (8). This database is not publicly available.

The criteria for inclusion are that the mutation is (i) reported in a peer-reviewed journal and (ii) assigned to a precise location in p53 DNA by DNA sequencing (of PCR-amplified material or cloned PCR products), or resequencing (e.g., using DNA microchip-based technologies). Mutations described in preliminary reports or in abstracts are not included. Data on loss of heterozygosity at the p53 locus, gene rearrangements and immunohistochemical analysis with anti-p53 antibodies are not part of this compilation. Also, mutations identified by digestion of DNA with restriction enzymes and demonstration of an RFLP are not entered. Mutations detected by screening techniques (such as SSCP, DGGE/CDGE or functional yeast assays) but not confirmed by DNA sequencing, are not included. Note that many papers initially retrieved by standard bibliographic search systems using 'p53' and 'mutation' as keywords will thus not figure in the reference list constituting the source of mutations in this database.

A typical entry in the database contains the description of one mutation found in a specified biological sample. The current version of the database contains 6598 entries, corresponding to 570 different mutational events. If identical samples and mutations were published in more than one article, only one of the reports is referenced and the data are entered only once in the database. If the identical mutation was found in two separate samples from the same patient, for example in primary tumor and in metastatic tissue or in cell line derived from the tumor, or in multiple primary tumors from the same patient, the mutation is entered as many times as it is associated with a different biological sample. However, if the same mutation is found in two separate biopsies at the same site and from the same patient, the mutation is presumed to be a single event and is entered only once.

Mutations identified in tumors are presumed to be somatic unless: (i) analysis of normal tissue from the same patient demonstrated that the mutation was constitutional in that individual; (ii) the mutation corresponded to one of the known constitutional polymorphisms of the human p53 gene (at codons 21, 31, 47, 72 and 213).

In many cases, the investigators have included screening techniques such as those mentioned above as a preliminary step for selecting material for sequencing. Information on these methodologies is not available in the current release of the database, but is being compiled and will be made available in later releases of the database.

STRUCTURE AND MANAGEMENT OF THE DATABASE

The database consists of a set of tables maintained as a relational database using Microsoft Access™. From this database, simple tables are produced in a flat-file format and are made publicly available in various electronic forms. Table A contains the description of each somatic mutation associated with a specified biological sample. This table contains 6598 entries (July 97 release). Table B lists the bibliographical records from which the entries in Table A are compiled, with title of the article, authors, abstract, key-words and cross-reference to Medline accession number (690 entries). Tables C and D will contain information on mutations in germline families and on polymorphisms in p53 exons, respectively. These two tables are presently being constructed and their full description will be available on the IARC p53 database website from January 1998 (see below). In the future, we intend to release additional tables containing further information (such as description of the methodology used to identify mutations or of known anamnestic, clinical or serological data).

The database is updated twice each year, in July and in January. The next release (January 1998) will coincide with the publication of the present article. This release will contain more than 8000 entries and will be numbered Release 1 (R1). Release 2 (R2) will be available in July 1998.

The database is managed in the Unit of Mechanisms of Carcinogenesis at IARC, under the supervision of a steering committee of IARC and external experts. Periodical review of the database content and formats by this steering committee ensures continuity in the development of the database as well as in the quality of the data. The development of this work is peer-reviewed by the IARC Scientific Council. Other p53 databases are publicly available in the scientific community, as well as sophisticated software for data analyses (9,10). However, the IARC database is the only one which is freely and entirely available in the form of raw flatfiles that can be exported in any computer format.

DESCRIPTION OF THE FLAT-FILE FORMAT

This section describes the format and content of Table A (somatic mutations). Each row (record) is a single mutation with arbitrarily assigned unique identity number.

Column 1

Unique mutation identity number. Tandem mutations (two adjacent base substitutions) are considered as one mutation event and are entered together, therefore tandem mutations have only one identity number and are a single record.

Column 2

An entry 'L' indicates that the material examined was from a tumor cell line. 'S' indicates that the material is from a surgical or autopsy sample (including fresh samples and archival, pathology specimen). 'B' indicates that the material is from a biopsy. 'X': other (such as xenograft) or unspecified.

Column 3

Nature of the mutation. The terms occurring in this column are 'Point' (missense and nonsense, and silent point mutations), 'insertion', 'deletion' and 'complex' (changes involving complex mechanisms such as substitution and insertion or deletion, or which are inherently ambiguous, e.g., base deletions in repetitive sequences) (see note below).

Column 4

Location of the mutation in the p53 gene. Terms occurring in this column are 'intron' (associated with intron number, 1–10) and 'exon' (associated with exon number, 2–11). Mutations in the promoter region and in the non-coding exon 1 are not included.

Column 5

For mutations in exons, codon number at which the mutation is located (1–393). If a tandem dinucleotide mutation spans two codons, both codons are entered. If other mutations span more than one codon, (e.g., deletion of several bases) only the first (5′) codon is entered. This information is not entered in the R1 version for complex mutations (see note below).

Column 6

For mutations in exons, nucleotide position at which the mutation is located in the p53 cDNA sequence (1–1179), numbered from the ATG codon to the termination codon. This information is not entered in the R1 version for complex mutations (see note below).

Column 7

For mutations in exons, normal base sequence of the codon in which the mutation occurred. This information is not entered in the R1 version for complex mutations (see note below).

Column 8

Mutated base sequence of the codon in which the mutation occurred. This information is not entered in the R1 version for complex mutations (see note below).

Column 9

Base change, read from the coding strand by convention, for base substitutions.

Column 10

For deletions and insertions, the number of bases deleted (–) or inserted (+) is given. For other mutations this column is left blank.

Column 11

Mutations that are single base transitions at CpG dinucleotides, i.e. CpG→TpG or CpG→CpA are designated by 'yes'. If the mutation does not fall into this category, 'no' is entered.

Column 12

Mutation type. This column assigned each mutation to a specific category in order to compile mutation spectra. Terms occurring in this column are 'GC:AT' (G→A and C→T transitions), 'GC:AT at CpG' (G→A and C→T transitions occurring at CpG dinucleotides), 'GC:TA' (G→T and C→A transversions), 'GC:CG' (G→C and C→G transversions), 'AT:TA' (A→T and T→A transversions), 'AT: GC' (A→G and T→C transitions), 'AT:CG' (A→C and T→G transversions), 'CC to TT' (double transitions at dipyrimidine repeats) and 'other' (other tandem mutations, insertions, deletions, complex mutations).

Column 13

Amino-acid encoded at the codon in which the mutation occurred (three letter amino acid abbreviation). No information is entered in the present version for complex mutations (see note below).

Column 14

Consequence of the mutation in the p53 protein sequence. For point mutations, the mutated amino-acid encoded at the codon in which the mutation occurred is indicated (three letter amino acid abbreviation). Chain terminating mutations due to single base substitutions are designated by 'stop'. Mutations that do not results in an amino acid change are designated 'silent'. Mutations that occurred in intron sequences are indicated by the term 'splicing', when they are suspected to result in splicing errors (however, it should be noted that in most instances these putative splicing errors were not verified; some of these base changes may be phenotypically silent). 'Frameshift': deletions, insertions and complex mutations suspected of producing a modification in the reading frame.

Column 15

The name or number given by the authors to the tumor sample or cell line is entered here. If the name is not distinctive, e.g., if the publication refers to samples as tumors 1,2,3, etc, then we have arbitrarily assigned a name, usually the first letters of the first author's name, followed by the numbers in the series.

Column 16

If more than one mutation has been found in the same sample, the unique identification number of corresponding entry (or entries) is given.

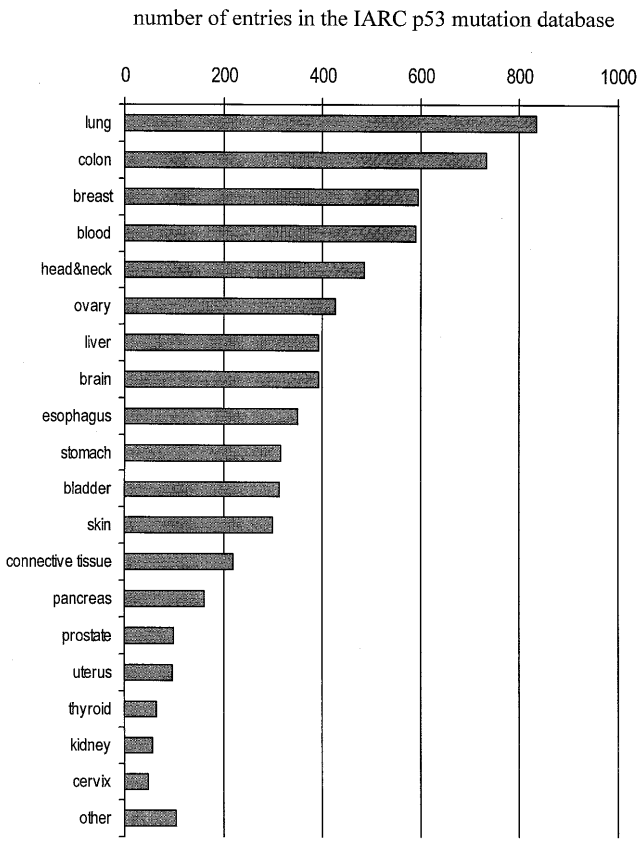

Figure 1. Distribution of tumors included in the IARC p53 mutation database.

Columns 17–20

These four columns are assigned to description of the tissue and type of lesion where the mutation has been identified. The heterogeneity of the terms occurring in columns 18 and 19 reflects the absence of common standards in the way tumors are described in the publications from which the data are compiled. We are currently modifying these entries to adopt a more standardized nomenclature method, whenever possible.

Column 17. Anatomic site, organ or tissue as described in the publication cited. The terms occurring in this column are listed in Figure 1.

Column 18. Tumor type as described in the publication cited. This column contains heterogeneous information. Examples of common terms used in this column are: ADC (adenocarcinoma); BCC (basal cell carcinoma); HCC (hepatocellular carcinoma); leukemia; lymphoma; melanoma; non-SCLC (non-small cell lung carcinoma); NPC (nasopharyngeal carcinoma); sarcoma; SCC (squamous cell carcinoma); SCLC (small cell lung carcinoma); TCC, (transitional cell carcinoma); Wilms (Wilm's tumors).

For abbreviations of sarcoma subtypes or leukemia and lymphoma subclassifications, e.g., ATL (adult T cell leukemia), refer to the cited publication. Uniformity of these abbreviations in the different reports has not been verified.

Column 19. Additional information on tumor or patient's history (such as tumor grade or stage and genetic predisposition). Examples of terms occurring in this column include: 'Barrett's' (Barrett's esophagus); FAP (familial polyposis coli); RER+ and RER– (replication error phenotype, positive or negative); Dukes A–D (Duke's stage A–D).

Column 20. ICDO classification number (International Classification of Diseases for Oncology, WHO) (11). When not available or ambiguous, this column is left blank.

Column 21

Reference number indicating the publication in which the mutation is described. The full citation (authors, title, journal, pages, year, abstract, key words, Medline identification number) is given as a separate table in Table B. This number corresponds to the unique identification number for each entry in Table B.

Columns 22–24

These three columns are assigned to the description of information on individual risk factors, exposures and ethnicity. They contain heterogeneous notes, usually comments emphasized by authors reporting the mutations. The diversity of the terms occurring in columns 22 and 23 reflects the heterogeneity of the information available in the publications from which these data are compiled. It should be noted that this information is generally qualitative. No quantitative information on exposure of risk factors is included in the database. This information does not presuppose that a formal, causal link has been established between such factors and the mutation described. Moreover, for most exogenous risk factors, individual exposure has not been monitored. This information is given solely to (i) permit the retrieval of mutations found in patients belonging to defined groups or having specific risk factors, and (ii) facilitate access to the corresponding publications. The data included in the database do not allow detailed comparison between exposure groups.

Column 22. Chemical, physical or biological risk factor associated with the mutation. The terms occurring in the column describe specific exposures [such as 'aflatoxin' (Aflatoxin B1); 'radon', 'asbestos', 'alcohol', 'dye' (occupational exposure in dye manufacturing); 'betel' (betel quid chewing), HPV (human papilloma virus); HBV (hepatitis B virus); EBV (Epstein–Barr Virus)]. Terms are preceded by '+' or '–', to indicate whether the mutation is considered to be associated or not with the specific exposure. When no information is available, this column is left blank.

Column 23. Information on sex and ethnicity. This information is entered only when specified in a publication as representing a possible factor influencing mutation specificity. Terms occurring in this column include 'male', 'female', 'Caucasian', 'Asian', 'African', 'American', 'European', as well as specific countries or geographic areas ['China'; 'Japan'; 'India'; 'Thailand'; 'Linxian' (Linxian Province, China), 'Normandy' (Normandy area, France)].

Column 24. Information on smoking status. Terms occurring in this column are 'smoker', 'nonsmoker' and 'exsmoker'. No data is provided on nature and quantity of tobacco consumption.

Column 25

When a record requires correction or information is added the date of the revision will be noted in this column. Record revisions are due to inherent inconsistencies or ambiguities in the published report or omissions or errors during data entry.

Example

The information on the entry No. 3966 is as follows:

Column 1: 3966: Unique mutation identity number.
Column 2: S: the sample analyzed is from surgery or autopsy.
Column 3: point: the mutation is a single base substitution.
Column 4: exon6: the mutation is localized in exon 6.
Column 5: 208: the mutation is located in codon 208 is 623.
Column 6: 623: the nucleotide position of the base change
Column 7: GAC: the wild-type codon sequence is GAC.
Column 8: GTG: the mutated allele sequence is GTG.
Column 9: A to T: the mutation is from A to T.
Column 10: (blank: the mutation is not an insertion/deletion).
Column 11: no: the mutation does not occur at CpG.
Column 12: AT:TA: the mutation is a transversion at an AT base pair.
Column 13: Asp: the wild-type amino-acid encoded at this codon is Aspartic acid.
Column 14: Val: the mutated amino-acid is valine.
Column 15: DG-6: the tumor sample name is DG-6.
Column 16: (blank: no other p53 mutation described in this sample).
Column 17: Lung: the anatomic site of the sample is lung.
Column 18: nonSCLC: the tumor is described as a non-small cell lung carcinoma.
Column 19: (blank: no additional information on tumor type, grade or stage).
Column 20: C34.9: The ICDO classification number is C34.9.
Column 21: 60: the mutation is reported in reference 60 in table D.
Column 22: +radon: the mutation had been found in a patient considered to have undergone exposure to radon.
Column 23: (blank: no specific data reported on sex or ethnicity).
Column 24: (blank: no information reported on smoking status).
Column 25: (blank: no corrections to the record since it was entered).

Reporting deletion, insertions and complex mutations

A HUGO initiative is underway to standardize and establish conventions for the nomenclature of mutations for specific locus mutation databases. We will attempt to follow the recommendations of this initiative in revising our database formats.

OBTAINING THE DATABASE

We consider that the p53 mutation database should be available as widely as possible and we see our task as a service to the scientific community. Ultimately, our objective is to provide access to the database in three distinct electronic formats, including (i) raw data in the form of flat-files, that can be modified and transferred in any electronic format according to user's needs; (ii) simple analysis programs allowing to compute basic operations such as search for mutational hotspots, analysis of mutation spectra and comparison of spectra between organ sites and tumor types; (iii) World Wide Web access for on-line analysis and visualisation of database contents. From January 1998, the IARC p53 website will provide access to the database in these three different electronic formats. To obtain the database, access the IARC p53 database website using the URL http://www.iarc.fr/p53/homepage.htm , selecting 'obtaining the database' and follow on-screen instructions. The flat-file format (see below) may also be obtained from EBI by one of the following methods: (i) anonymous ftp to ftp.ebi.ac.uk, in the directory/pub/database/p53; (ii) World Wide Web access using the URL http://www.ebi.ac.uk/ , selecting 'services'going to the databases selection and selecting 'IARC p53'.

Flat-file

Table A (somatic mutations) and Table B (bibliographic list) are available as Excel™ spreadsheets, which requires use of the Microsoft Excel™ program on either an MS-DOS™ or Windows™ systems or an Apple Macintosh™. The data has also been converted into a flatfile format modeled after the standard used by the EMBL nucleotide sequence database. In this format the data are stored in an ASCII text file with each column of the spreadsheet represented by a special line type. The flatfile format can be used on any computer system and with standard text editors. The p53 directory contains the original spreadsheet file as a Macintosh binHex4.0 self-extracting archive. The release notes are included in the file p53.doc and Table B is in p53.ref. Also included is the database in flatfile format (p53.dat) and the data in tab-delimited (data.tab) and comma-delimited (data.comma) formats for usage by other data management systems.

Analysis system under Claris FileMakerPro 3.0™

Using the data in Tables A and B, we have developed a semi-relational database under FileMakerPro 3.0™. This requires the use Claris FileMakerPro 3.0™ software on Apple Macintosh™ or under Windows95™ or WindowsNT™ (or higher). In this version of the database, each of the 25 columns described above represents a searcheable, indexed field. This program allows the retrieval, sorting and analysis of the data on the basis of any specific query corresponding to one or several term(s) or number(s) occurring in the database. Different layouts (individual records, tables, summary tables) have been developed to visualise the results of database searches. Build-in calculation routines allow for easily performed operations such as searching for mutational hotspots or determining mutation spectrum in a dataset. The database also contains indexed bibliographic records that can be searched on the basis of any term occurring in title, author names or abstract. This version of the database is simple to use and is a powerful tool for answering most of the questions commonly addressed to the database.

To access this version of the database, it is necessary to acquire the FileMakerPro 3.0™ software. Our database does not run on earlier versions such as FileMakerPro 2.0™. Please note that a minimal computer configuration equivalent to Pentium 100 MHz with 16 Mb RAM is required for efficient running of the program. At present, Claris is making available a time-limited, free trial version of their new software FileMaker Pro 4.0beta™ (at the URL http://www.filemaker.com/beta/newsrelease.html). This new version supports our database and enables the production of

dynamic web pages. We will use this software to give direct, on-line access to our database in 1998.

World-Wide-Web access

The IARC p53 Website (at the URL http://www.iarc.fr/p53/homepage.htm) provides access to a comprehensive source of information on p53 mutations and the database. This includes: (i) review of the most significant lessons from p53 mutation analysis; (ii) description of the database structure, contents and management; (iii) index of all the searchable terms occurring in the database; (iv) access to different versions of the database; (v) static pages with graphic results of p53 mutation analysis in all common tumors, classified by organ site and by tumor type; and (vi) links with other, related databases (such as the MDM2 database) and with other websites containing p53 or related genetic information.

At EBI, the flat-file version of the database has been indexed using the SRS 5.0 (Sequence Retrieval System) (12). This system allows the user to query the database in order to retrieve and download subsets of records matching specific criteria. For more information and access to SRS 5.0, follow the links on the EBI server.

THE EBI MUTATION VIEWER: A NEW TOOL FOR VISUALISATION OF THE IMPACT OF MUTATIONS ON p53 PROTEIN STRUCTURE

The development of new visualisation tools is essential for better exploitation of the information on p53 mutations. One important aspect of the analysis of p53 mutations is to understand how these mutations affect the structure and function of the p53 protein. Portions of the p53 proteins have been crystallized and several structures are now published, including that of the DNA-binding domain of the protein (residues 102–292, containing most of the mutations associated with cancer) (13). In order to locate the mutations described in the database in the structure of the p53 protein, a Mutation Viewer Java applet has been designed by Alan Robinson at EBI (Fig. 2). This applet consists of classes written in Java that makes use of the IARC p53 database. It allows the user to query the database and to visualise the data as four integrated windows, including: (i) a view of the primary structure of the protein showing conserved and functional domains; (ii) a representation of the topographical relationships between secondary structure elements; (iii) a dynamic, three-dimensional representation of the p53-DNA co-crystal (13); and (iv) a text message window. The p53 database has been chosen as an example database that could be used with this applet, which may be adapted to databases of other genes. The applet is available from EBI and from the IARC p53 website.

The down-loading of the applet involves two stages, (i) the down-loading of the viewer classes after which four windows appear, and (ii) the down-loading of the database from EBI in the flat-file format. These operations may take between 3 and 5 min on a PC. Once this is done, a window and query box appear in the browser window. When search terms are entered in the query box, the results are highlighted as 'hits' in the different representations of the protein. When point hit is selected in any window, it is highlighted in all the display, and the entries at the corresponding codon position are shown in the text message window. Another tool to visualize p53 mutations has been developed by Moeckel *et al*. (14) using the Virtual Reality Modelling Language and is also available on the World Wide Web (http://pc.chemie.th-darmstadt.de/vrml/p53dna).

RECOMMENDATIONS TO AUTHORS AND USERS OF THE DATABASE

The development of the database relies exclusively on the quality and accuracy of published records. In recent months, we have noticed a trend in the literature to publish data from mutation analysis in the form of summary graphs or tables, while the detailed, individual mutation data are not reported. When the information is not given in the publication, we recommend that authors submit the full list of mutations to the IARC p53 database at p53database@iarc.fr. In the future, we will develop direct data submission forms to ensure that this valuable information is not lost for the scientific community. Since January 1997, our policy has been to contact the authors of all publications containing incomplete or ambiguous records to ask for complementary information.

When tumor mutations are reported for the second time in a new publication we recommend this be stated in a footnote to the table where the mutations are re-listed, also indicating which tumor mutations were reported previously. Providing tumor samples with unique case numbers would also help to avoid redundancies in the database.

Users of the database should be aware that the data in the database cannot be considered as representing an exhaustive description of the frequency, nature and distribution of p53 mutations in human cancers. Most of the data included in the database are from studies where tumor cases are selected on the basis of relevance to pathological or clinical questions. Only a minority of the publications are molecular epidemiological studies with adequate controls and exposure groups. Another, important bias is that many studies have analyzed only exons 5–8, as this region has been shown to contain the majority of the mutations associated with human cancer. It is recommended that database users refer to original publications for detailed interpretation of mutation data.

p53 MUTATIONS AND HUMAN CANCER

The diversity of p53 mutations is useful to identify possible sources of cancer-causing mutation in the human setting. Mutagens and carcinogens damage the genome in characteristics ways, leaving 'mutagen fingerprints' in DNA. Specific DNA changes can also arise from endogenous biological processes. However, DNA-repair processes and bioselection of mutants with specific properties act as additional 'filters' in determining which mutations will be stabilized and retained during the progression of cancer. How specific targeting and bioselection both contribute to tumor-specific mutation spectra has been extensively discussed in several recent reviews (4,5,15).

The p53 gene is located on chromosome 17p13 and comprises 11 exons within a chromosomal domain of ~20 kb. The p53 protein is a multi-functional transcription factor which plays a role in the control cell cycle progression, DNA integrity and cell survival in cells exposed to DNA-damaging agents. DNA damage induces a transient nuclear accumulation and activation of p53, resulting in the transactivation of target genes such as the cyclin kinase inhibitor $p21^{waf-1}$ and the regulator of apoptosis

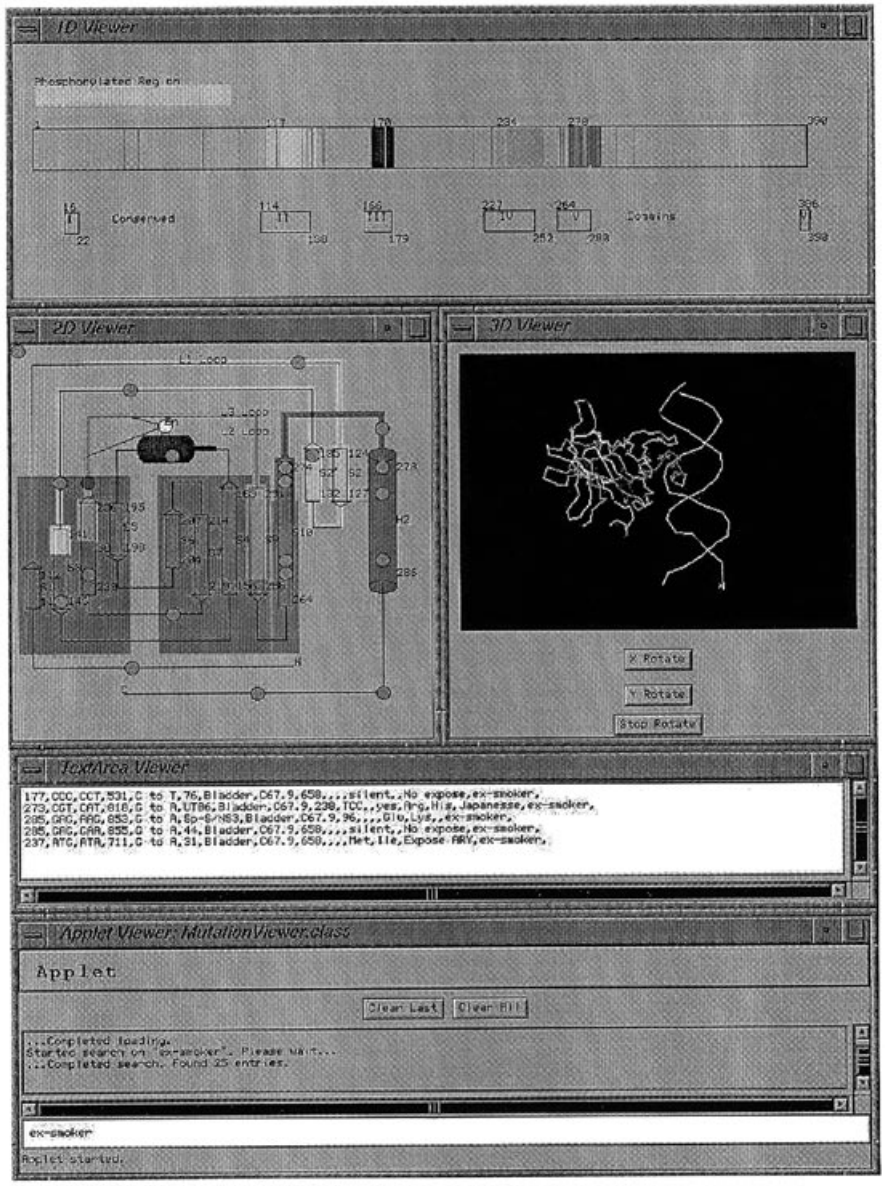

Figure 2. The EBI mutation viewer applet. This applet includes three graphic windows displaying pictures of the p53 protein (1D, 2D and 3D viewers), a text area viewer and a query box (applet viewer). Data from the IARC p53 mutation database are visualised as 'hits' in the three graphic windows and the description of the corresponding database entries are given in the text area. For further explanations, see text.

bax-1. The p53 protein also interacts with several proteins regulating DNA replication, transcription and repair, and can exert suppressive effects by transcription-independent pathways. Most mutations associated impair the transactivating properties of p53 and prevent growth arrest or apoptosis in response to DNA-damage. In addition, some mutants may exert an oncogenic activity of their own. The molecular basis for this 'gain-of-function' phenotype is still unclear (1–3).

p53 mutations as clues to the etiopathogenesis of cancer

Mutations appear as the result of endogenous processes or of the action of exogenous, physical or chemical carcinogens. Endogenous processes include methylation and deamination of cytosine at CpG residues, injury by free radicals generated by biological processes, and errors in DNA-repair and synthesis. Moreover, mutations do not form at an equal rate at all base positions. For example, formation of adducts in p53 DNA by metabolites of benzo(*a*)pyrene occurs preferentially on guanines adjacent to methylated cytosine. DNA repair is also thought to be more efficient in actively transcribed genes than in silent DNA regions, and to be faster on the DNA strand used as a template for transcription than on the coding strand. Thus, a strand bias in the distribution of mutations may be indicative of the involvement of exogenous carcinogens, although experimental evidence for the role of transcription-coupled repair in such bias is still inconclusive (15).

Early analyses of the database have revealed examples of mutation patterns consistent with fingerprints of DNA damage induced by defined exogenous carcinogens. These examples

include G→T transversion at codon 249 (AGG→AGT) in hepatocellular carcinomas of patients with high dietary exposure to Aflatoxin B1 (reviewed in ref. 16) and tandem CC→TT transitions, a typical signature of UV-induced mutagenesis, in non-melanoma skin cancer (17). However, in most cancers, the mutation pattern is complex. Cancers associated with tobacco smoking are a good illustration of this complexity. Smoking is a major risk factor in several cancers, including cancers of the oral cavity, esophagus, lung and bladder. In all pathologies, the prevalence of p53 mutations is generally higher in smokers than in non-smokers (18–20). However, the spectrum of p53 mutations varies from one pathology to the other. In lung cancers, G→T transversions, a typical signature of mutagenesis by benzo(*a*)pyrene, are particularly frequent. Experimental evidence shows that benzo(*a*)pyrene preferentially forms adducts at codons 157, 248 and 273, which are all mutation hotspots in lung cancers (21). In squamous cell carcinoma of the oral cavity and of the esophagus, the predominant types of mutations are transitions or transversions at A:T base pairs and G→A transitions. In these cancers, the combined consumption of tobacco and alcohol is considered to be a cumulative risk factor, and the observed spectrum is consistent with a role of *N*-nitrosamines (G:C transitions) and of metabolites of ethanol, such as acetaldehyde (mutations at A:T bases) (18). In bladder cancer, the mutation spectrum is dominated by mutations at G:C base pairs (75%, including 28% of transitions at non-CpG). The major tobacco carcinogen(s) in these tumors are thought to be aromatic amines such as 4-aminobiphenyl. Interestingly, the mutation spectrum is similar in bladder cancers of workers occupationally exposed to aromatic amines as in the general population, supporting the notion that the carcinogens involved are identical in both groups (22,23). Thus, a complex carcinogenic mixture such as tobacco smoke can have a different mutagenic impact in different tissues. These examples demonstrate that more sophisticated molecular epidemiological studies with exposure cohorts matched for various parameters that could influence the mutation spectrum (such as age, sex, ethnic origin, etc.) are required to determine the role of exogenous agents in the generation of such mutation spectra.

Functional implications of p53 mutations

A large proportion of the literature on p53 mutations addresses the usefulness of mutation analysis in the molecular pathology of cancer. Mutations may serve as molecular indicators of clonality or as early markers of relapse in a patient with a previously identified mutation in the primary tumor. As p53 plays an essential role in the cellular response to DNA-damage, the mutation status may be an important determinant of the tumor response to chemo- or radiotherapy. Detection of antibodies to p53 in the serum of cancer patients may provide an interesting tool for diagnosis and follow-up of cancer (24).

Analysis of the p53 mutation database reveals that 95% of known mutations fall within the DNA-binding domain (see above). The DNA-binding domain is made of two antiparallel β-sheets that form a scaffold supporting a DNA-binding surface made of non-contiguous loops and helixes, stabilized by an atom of zinc (13). Many of the residues commonly mutated in cancer3 are involved in direct contacts between p53 and DNA (such as arginines 248, 273 and 282). Other common hotspot residues such as arginine 175 and 249 are not in direct contact with DNA but their mutation may alter the architecture of the DNA-binding surface. There is evidence that individual mutants differ by their biological properties in experimental systems (25) and by their consequences for the progression and prognosis of cancer (26,27). Moreover, the functional properties of mutant p53 may also be cell-type specific (28). Analysis of the database indicates that the distribution of mutations within the structure of p53 varies from one tumor type to another. However, more detailed analysis is required to determine whether different tumor types select for mutants with different functional properties.

FUTURE PROSPECTS

In the past, the p53 mutation database has proven a valuable tool for molecular biologists, allowing the generation of hypotheses that could be tested in the laboratory. The accumulation of data on p53 mutations is now generating high expectations in molecular pathology and molecular epidemiology. However, exploitation of p53 mutation data in these fields will require the solving of a number of practical questions. Firstly, an important effort has to be made to establish recognized standards for study design, choice of methods, quality control and interpretation of results. Secondly, it is essential that detailed and specific information is reported on the techniques used for mutation analysis, on the identification of the tumor pathology and grade, on the individual characteristics of the tumors examined (age and ethnicity of the patients, prognosis, response to treatment), and on individual exposures to cancer risk factors. The compilation of this information in powerful databases containing high quality, peer-reviewed data, is a major challenge for the years to come.

ACKNOWLEDGEMENTS

We thank Mrs M. Wrisez for secretarial assistance. This work is supported in part by EU Contract BI04-CT96-0346.

REFERENCES

1 Ko, L.J. and Prives, C. (1996) *Genes Dev.*, **10**, 1054–1072.
2 Harris, C.C. (1996) *Carcinogenesis*, **17**, 1187–1198.
3 Levine, A. (1997) *Cell*, **89**, 323–331.
4 Lehman, T.A. and Harris, C.C. (1994) *IARC Sci. Publ.*, **125**, 399–412.
5 Greenblatt, M.S., Bennett, W.P., Hollstein, M. and Harris, C.C. (1994) *Cancer Res.*, **54**, 4855–4878.
6 Harris, C.C. (1996) *Br. J. Cancer*, **73**, 261–269.
7 Hollstein, M., Rice, K., Greenblatt, M.S., Soussi, T., Fuchs, R., Sorlie, T., Hovig, E., Smith-Sorensen, B., Montesano, R., and Harris, C.C. (1994). *Nucleic Acids Res.*, **22**, 3551–3555.
8 Kleihues, P., Schauble, B., zur-Hausen, A., Esteve, J. and Ohgaki, H. (1997) *Am. J. Pathol.*, **150**, 1–13.
9 Cariello, N.F., Douglas, G.R and Soussi, T. (1996) *Nucleic Acids Res.*, **24**, 119–120 [see also this issue, *Nucleic Acids Res.* (1998) **26**, 198–199].
10 Beroud, C., Verdier, F. and Soussi, T.(1996) *Nucleic Acids Res.*, **24**, 147–150 [see also this issue, *Nucleic Acids Res.* (1998) **26**, 200–204].
11 World Health Organization (1990) In International Classification of Diseases of Oncology. Percy, C., Van Hollens, V. and Muir, C. (eds), World Health Organization, Geneva, 200–204.
12 Etzold, T. and Argos, P (1993) *Comput. Appl. Biosci.*, **9**, 49–57.
13 Cho, Y., Gorina, S., Jeffrey, P.D. and Pavletich, N.P. (1994) *Science*, **265**, 346–355.
14 Moeckel, G., Keil, M., Hollstein, M., Spiegelhalder, B., Bartsch, H. and Brickmann, J. (1997) *J. Mol. Model.*, **3**, 382–385.
15 Dogliotti, E. (1996) *Carcinogenesis*, **17**, 2113–2118.
16 Montesano, R.M., Hainaut, P. and Wild, C. (1998) *J. Natl. Cancer Inst.*, In press.

17 Nakazawa, H., English, D., Randell, P.L., Nakazawa, K., Martel, N., Armstrong, B.K. and Yamasaki, H. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 360–364.
18 Montesano, R., Hollstein, M. and Hainaut, P. (1996) *Int. J. Cancer*, **69**, 225–235.
19 Brennan, J.A, Boyle, J.O, Koch, W.M, Goodman, S.N., Hruban, R.H., Eby, Y.J., Couch, M.J., Forastiere, A.A. and Sidransky, D. (1995) *N. Engl. J. Med*., **332**, 712–717.
20 Kondo, K., Tsuzuki, H., Sasa, M., Sumitomo, M., Uyama,T. and Monden, Y. A (1996) *J. Surg. Oncol*., **61**, 20–26.
21 Denissenko, M.F., Pao, A, Tang, M. and Pfeifer, G.P. (1996) *Science*, **274**, 430–432.
22 Taylor, J.A., Li, Y., He, M., Mason, T., Mettlin, C., Vogler, W.J., Maygarden, S. and Liu, E. (1996) *Cancer Res*., **56**, 294–298.
23 Sørlie, T., Martel-Planche, G., Hainaut, P., Lewalter, J., Holm, R., Borresen-Dale, A.L. and Montesano, R. (1998) *Br. J. Cancer*, In press.
24 Sidransky, D. and Hollstein, M. (1996) *Annu. Rev. Med.* **47**, 285–301.
25 Ory, K., Legros, Y., Auguin, C. and Soussi, T. (1994) *EMBO J*., **13**, 3496–3504.
26 Goh, H.S., Yao, J. and Smith, D.R. (1995) *Cancer Res*., **55**, 5217–5221.
27 Aas, T., Børresen, A.L., Geisler, S., Smith-Sorensen, B., Johnsen, H., Varhaug, J.E., Akslen, L.A. and Lonning, P.E. (1996) *Nature Med*., **2**, 811–814.
28 Forrester, K., Lupold, S.E., Ott, V.L., Chay, C.H., Band, V., Wang, X.W. and Harris, C.C. (1995) *Oncogene*, **10**, 2103–2111.