200–204 *Nucleic Acids Research, 1998, Vol. 26, No. 1*

p53 gene mutation: software and database

Christophe Béroud and Thierry Soussi[1,*]

Hôpital Necker Enfants Malades, U383 INSERM, Paris, France and [1]UMR 218 du CNRS, Institut Curie, Pavillon Trouillet Rossignol, 26 rue d'Ulm, 75231 Paris cedex 05, France

Received October 14, 1997; Accepted October 15, 1997

ABSTRACT

A large number of different mutations in the p53 tumor suppressor gene have been identified in all types of cancer. As of October, 1997, this database (http://perso.curie.fr/tsoussi) contained >7500 mutations. Such a substantial increase since our previous reports should enable epidemiological analyses which were not previously possible. In order to analyse these new data, the UMD software has been improved. A new Web version of the UMD software enables online analysis of the database. The present report describes various improvements since the last release of the database.

INTRODUCTION

We had previously described generic software that enables the entry and analysis of mutations in any gene of interest. This software was used for the creation and analysis of various mutation databases that have been described over the past few years: the p53 gene (1), the APC gene (2), the fibrillin gene (3), the VHL gene (C. Béroud *et al*., unpublished results), the WT1 gene (C. Béroud *et al*., unpublished results) and the LDLR gene (4).

The p53 database was set up in 1991 and the first publication of the database in 1992 contained 300 mutations (5). The latest release of the p53 database contains 7434 mutations taken from 820 articles (October 1997) (Table 1). Since the 1997 release of the p53 database, a large number of changes have come about both in the database itself and in the software used to manage it (1,6,7).

THE 1998 VERSION OF THE p53 DATABASE

Systematic analysis of the old version of the database revealed that there were a number of duplicate entries due to multiple publications of identical cases. Duplications were checked by searching the database using three criteria: name of the patient, tumor type and authors. Clear-cut duplications were deleted and only the first publication was included. When in doubt, the authors were contacted in order to clarify the situation. More than 200 mutations were deleted from the 1997 p53 database following this analysis. After taking these precautions, we believe that <1% of the data is now duplicated.

UMD software

Several new features have been added to the software. The structure of the central domain of the p53 protein has been elucidated leading to the discovery that several regions of the p53 protein are essential for its DNA binding property (8). More recent studies have suggested that there is substantial heterogeneity in the behavior of the various p53 mutants (9–11). Such heterogeneity has been linked to a sharp difference in the clinical response (12–14). According to the X-ray structure of p53, each position in the central region of the p53 protein has been linked to structural data.

New features have been added to the software in order to analyze the etiology of skin tumors in relation to UV exposure (15–17). Since the most recent release of the p53 database, the number of mutations in various types of skin cancer has increased by 100%. Such mutations have a very peculiar pattern in relation to the lesion caused by the UV (Fig. 1). UV-radiation-induced mutations have been studied in various animal models. Most of the mutations are found to be located at dipyrimidine sites (i.e., T-T, C-C, C-T or T-C) and correspond to a C→T transition. More than 20% correspond to tandem mutations involving the two adjacent nucleotides of the dipyrimidine sites (C-C→T-T). Xeroderma pigmentosum (XP) is an autosomal, recessively inherited disease. Patients with XP show clinical and cellular hypersensitivity to UV radiation, resulting in a very high incidence of skin cancer. Analyses of 325 skin cancer mutations revealed important informations (Fig. 1 and Table 2). Most mutational events leading to p53 inactivation in skin cancer are GC→CT transitions, as in colon cancer. Nevertheless there are some significant differences between these two cancers; (i) transitions in colon cancer and in other internal cancers are found equally at dipyrimidine and non-dipyrimidine sites whereas a majority of transitions found in various skin cancers are predominantly found at dipyrimidine sites (Fig. 1A), which are the primary sites for UV photo lesions; (ii) most transitions observed in colon cancer are G→A or C→T modifications that are localized at CpG dinucleotides. It is well known that spontaneous deamination of 5-methylcytosine at these nucleotides may be an important cause of this type of transition. In skin cancer, most transitions are C→T localized in the non transcribed strand of the DNA. Transversions, such as GC→TA, which are predominant in lung cancer, are not found in skin cancer (Table 2).

*To whom correspondence should be addressed. Tel: +33 1 42 34 65 11; Fax: +33 1 42 34 67 25; Email: thierry.soussi@curie.fr

Table 1. The 1998 issue of the p53 database

	No. of entries	Point mutations missense/nonsense	Frameshift mutations deletion/insertion
Entire p53 database	7434	5987/530	673/169
Somatic mutation database	6613	5316/478	608/148
Germline mutations	110	93/6	8/3
Cell lines mutations	655	532/44	55/17

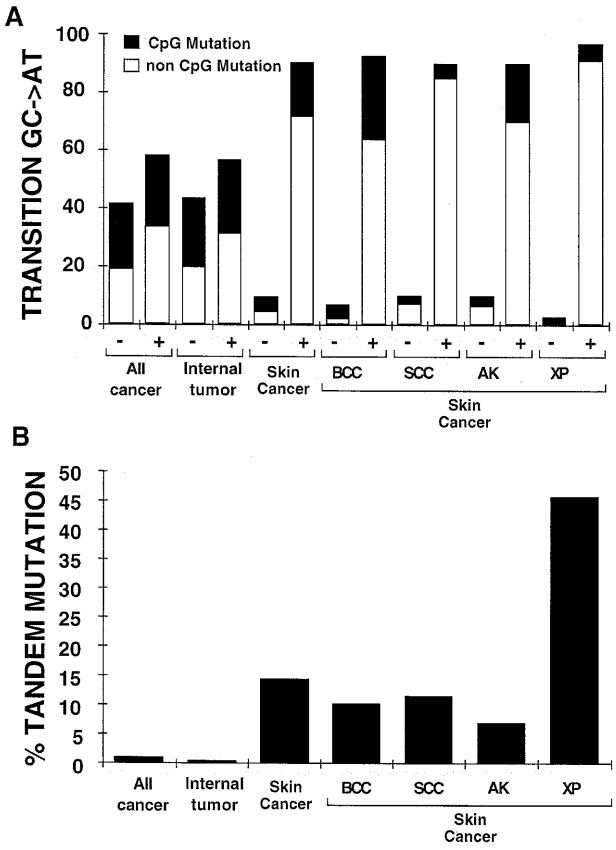

Figure 1. (**A**) Distribution of GC→AT transitions at Py-Py (+) and non Py-Py (–) sites. The white and black columns correspond to mutations at non CpG and CpG sites respectively. BCC, basal cell carcinoma; SCC, squamous cell carcinoma of the skin; AK, actinic keratosis; XP, xeroderma pigmentosum patients. (**B**) Distribution of tandem mutations in various cancers types. Abbreviations are described in (A).

Although the p53 gene is usually inactivated by missense mutations, 10–15% of the mutations are small frameshift mutations (insertions or deletions). An exhaustive analysis of these frameshift mutations has been performed recently (18). Most of them do not occur at random position as they are found at specific DNA sequences which are known to be associated with DNA polymerase infidelity: monotonic base runs, adjacent or non-adjacent repeats of short tandem sequences, palindromes or homocopolymer runs. A new feature to analyse DNA region around deletion and insertion was added to the software (Fig. 2). It allows the detection of repeated sequence which are adjacent to the mutations.

Table 2. Comparison of mutational events between internal and skin cancer

	ALL	Internal	Colon	Skin	Lung
G->A	11.9	11.9	12.1	12.9	9.0
G->A at CpG	15.9	16.4	**29.4**	4.2	5.3
C->T	8.7	7.6	4.3	**31.8**	8.2
C->T at CpG	13.7	13.4	**24.9**	19.2	7.1
total	50.2	49.3	70.7	68.1	29.6
A->G	9.3	9.7	5.2	1.4	9.4
T->C	3.6	3.6	2.6	3.5	1.5
A->T	3.2	3.1	2.1	3.5	3.2
A->C	1.5	1.5	0.8	1.4	1.7
T->G	2.8	2.8	1.1	2.8	2.2
T->A	3.1	3.2	2.8	1.0	2.4
C->A	2.6	2.5	2.1	3.5	2.6
C->G	3.5	3.5	1.2	3.1	4.4
G->T	15.0	15.4	8.6	7.0	**34.6**
G->C	5.2	5.2	2.9	4.5	8.6
nb cases	6424	6138	654	286	723

Only missense and nonsense mutations were used in this analysis.

The p53 Web site

A new p53 Web site was created in 1997 (Fig. 3). This site contains extensive information on various aspects on p53 and also links to other p53 sites. The description of all the different features of this site is beyond the scope of this review, but the most important aspect is that we have developed several specific pages dedicated to molecular epidemiology studies. A p53 mutation analysis page offers the opportunity to study and download specific informations on a given type of cancer (Fig. 4). For each cancer type, a file of p53 mutations specific for this cancer can be downloaded. Other useful features can be found by navigating through this site.

The Web version of UMD software

One important novelty of the software is the ability to perform analyses directly on the Web. The UMD sofware is now available online on the web with various features for the analysis of the p53 database. The most important feature is called ‘New mutation?’. An essential question following the identification of a mutation in a new tumor sample concerns the occurrence of such mutation in the database. Does the mutation occur at a hot spot codon (if indeed it is a true mutation) or is this a position that has never been found mutated so far suggesting that one should verify again the

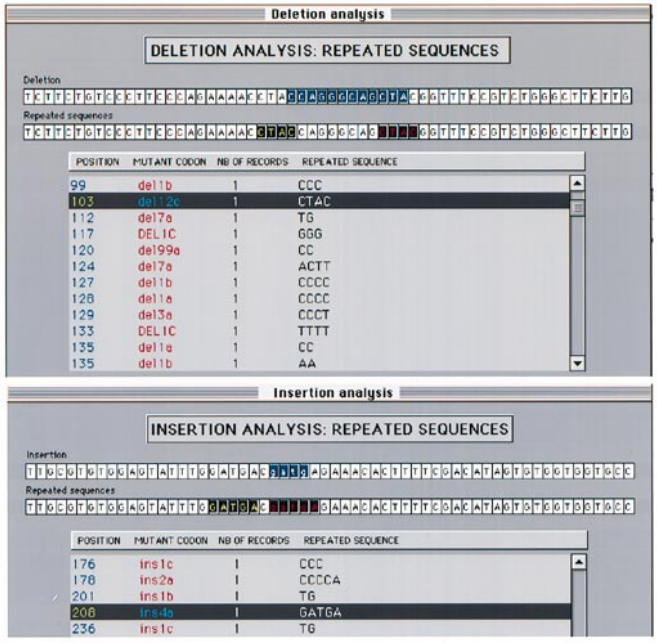

Figure 2. Deletion and insertion analysis. About 12% of p53 mutations are caused by frameshift mutations. More than 80% are localized in region with tandem repeats, monotonic base runs or palindromes (18).

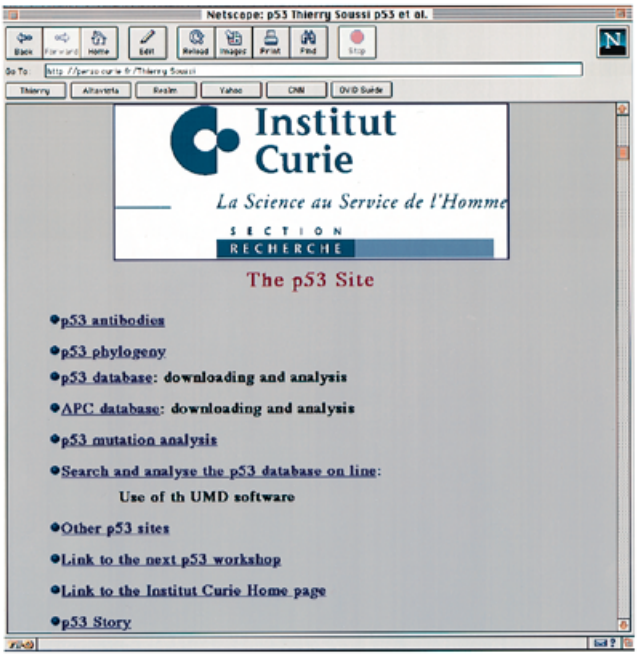

Figure 3. Home page of the p53 site localized at the Institut Curie (Paris, France). The p53 databases described in Table 1 can be downloaded from this site.

sequence of the sample (Fig. 5)? In addition, other analyses tools and listings are available.

Availability

The program was developed with the 4th Dimension (4D) package from ACI (version 6.0). This version generates a compiled program which can be used either on a MAC (680xx or PowerMac) or an IBM computer (486 or Pentium). Furthermore, the compilation integrates the runtime of 4th Dimension, alleviating the need for any other software. This stand-alone software has a size of 9 Mb. The p53 database has a size of 10 Mb. The software and the databases are freely available. They can be requested from T. Soussi (thierry.soussi@curie.fr) or C. Béroud (béroud@ceylan.necker.fr). Ten formatted floppy disks are necessary for sending the full version of the database and the software. A solid support such as a Syquest cartridge (44 Mb, 88 Mb or 200 Mb), Jaz cartridge (1 Gb) or a CD can also be handled.

The p53 database is also available on our Web site at the following address: http://perso.curie.fr/tsoussi . Four p53 databases are available (Table 1): the first contains all published p53 mutations; the second contains only somatic mutations found in human tumors; the third contains germline mutations found in cancer families; the fourth contains mutations found in cell lines. Each file can be downloaded from our site. They have been saved as Microsoft Excel files and can be used either with a Mac or an IBM. All databases are also available by mail (two formatted floppy disks are necessary). The server with the Web version of the UMD software is available at the following address: http://perso.curie.fr/tsoussi

ACKNOWLEDGEMENTS

We thank J. Bram for critically reading this manuscript. This work was supported by grants from the Association de Recherche sur le Cancer, the Ligue Nationale contre le Cancer (Comité National and Comité de Paris) the MGEN and the Fondation pour la Recherche Médicale.

REFERENCES

1 Béroud, C., Verdier, F. and Soussi, T. (1996) *Nucleic Acids Res.*, **24**, 147–150.
2 Béroud, C. and Soussi, T. (1996) *Nucleic Acids Res.*, **24**, 121–124 [see also this issue, *Nucleic Acids Res.* (1998) **26**, 269–270].
3 Collod, G., Béroud, C., Soussi, T., Junien, C. and Boileau, C. (1996) *Nucleic Acids Res.*, **24**, 137–140.
4 Varret, M., Rabes, J. P., Collod-béroud, G., Junien, C., Boileau, C. and Béroud, C. (1997) *Nucleic Acids Res.*, **25**, 172–180 [see also this issue, *Nucleic Acids Res.* (1998) **26**, 248–252].
5 Caron de Fromentel, C. and Soussi, T. (1992) *Genes Chrom. Cancer*, **4**, 1–15.
6 Cariello, N. F., Cui, L., Béroud, C. and Soussi, T. (1994) *Cancer Res.*, **54**, 4454–4460.
7 Béroud, C. and Soussi, T. (1997) *Nucleic Acids Res.*, **25**, 138–138 [see also this issue, *Nucleic Acids Res.* (1998) **26**, 269–270].
8 Cho, Y. J., Gorina, S., Jeffrey, P. D. and Pavletich, N. P. (1994) *Science*, **265**, 346–355.
9 Ory, K., Legros, Y., Auguin, C. and Soussi, T. (1994) *EMBO J.*, **13**, 3496–3504.
10 Rowan, S., Ludwig, R. L., Haupt, Y., Bates, S., Lu, X., Oren, M. and Vousden, K. H. (1996) *EMBO J.*, **15**, 827–838.
11 Friedlander, P., Haupt, Y., Prives, C. and Oren, M. (1996) *Mol. Cell. Biol.*, **16**, 4961–4971.
12 Goh, H.S., Yao, J. and Smith, D.R. (1995) *Cancer Res.*, **55**, 5217–5221.
13 Bergh, J., Norberg, T., Sjogren, S., Lindgren, A. and Holmberg, L. (1995) *Nature Med.*, **1**, 1029–1034.
14 Borresen, A.L., Andersen, T.I., Eyfjord, J.E., Cornelis, R.S., Thorlacius, S., Borg, A., Johansson, U., Theillet, C., Scherneck, S., Hartman, S., Cornelisse, C.J., Hovig, E. and Devilee, P. (1995) *Genes Chromosomes Cancer*, **14**, 71–75.
15 Brash, D.E., Rudolph, J.A., Simon, J.A., Lin, A., Mckenna, G.J., Baden, H.P., Halperin, A.J. and Ponten, J. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 10124–10128.

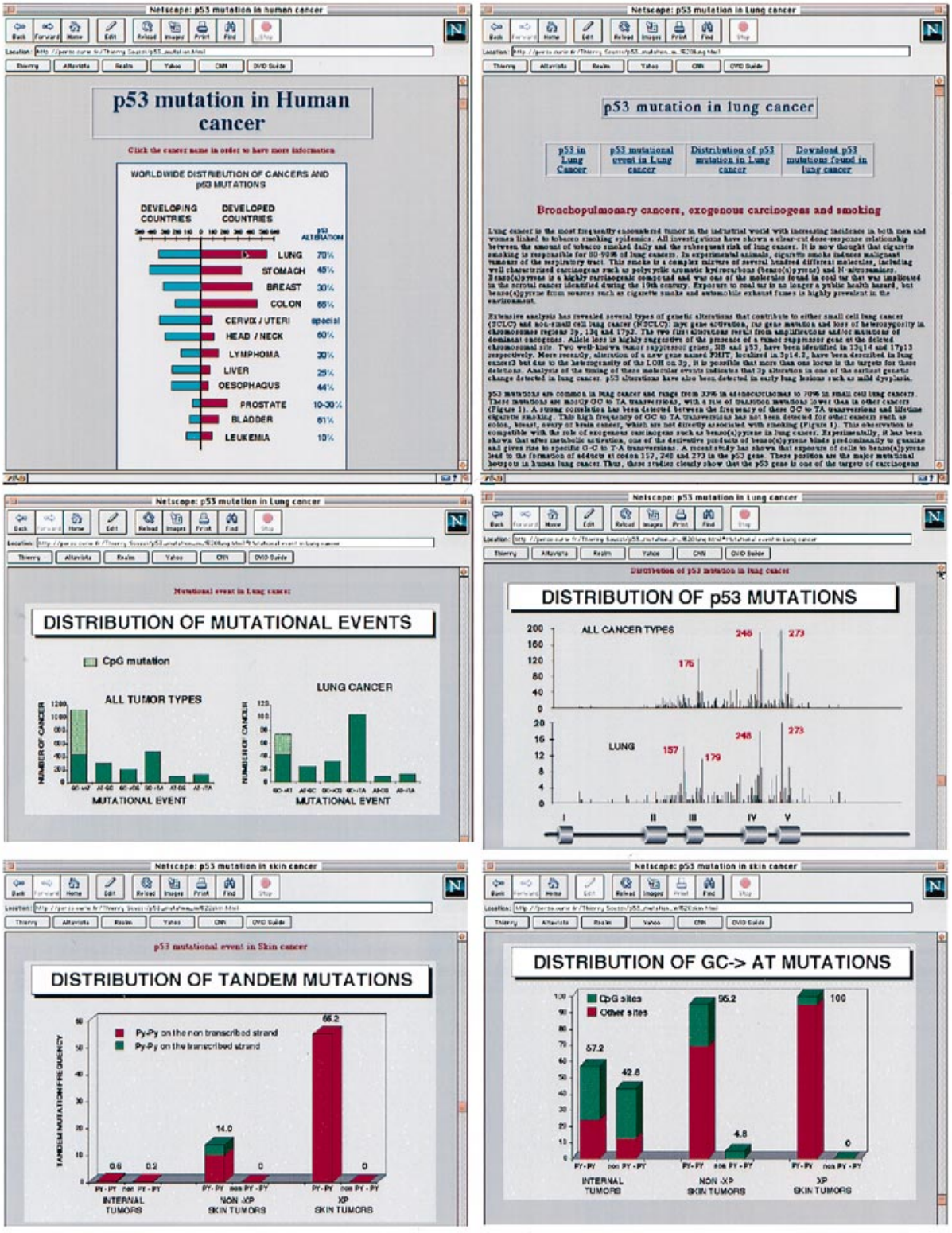

Figure 4. p53 mutation analysis. For various types of cancer there are links to specific pages with epidemiological data and possibilities for downloading specific subsets of the p53 database.

16 Dumaz, N., Stary, A., Soussi, T., Dayagrosjean, L. and Sarasin, A. (1994) *Mutat. Res.*, **307**, 375–386.

17 Ponten, F., Berg, C., Ahmadian, A., Ren, Z.P., Nister, M., Lundeberg, J., Uhlen, M. and Ponten, J. (1997) *Oncogene*, **15**, 1059–1067.

18 Greenblatt, M.S., Grollman, A.P. and Harris, C.C. (1996) *Cancer Res.*, **56**, 2130–2136.

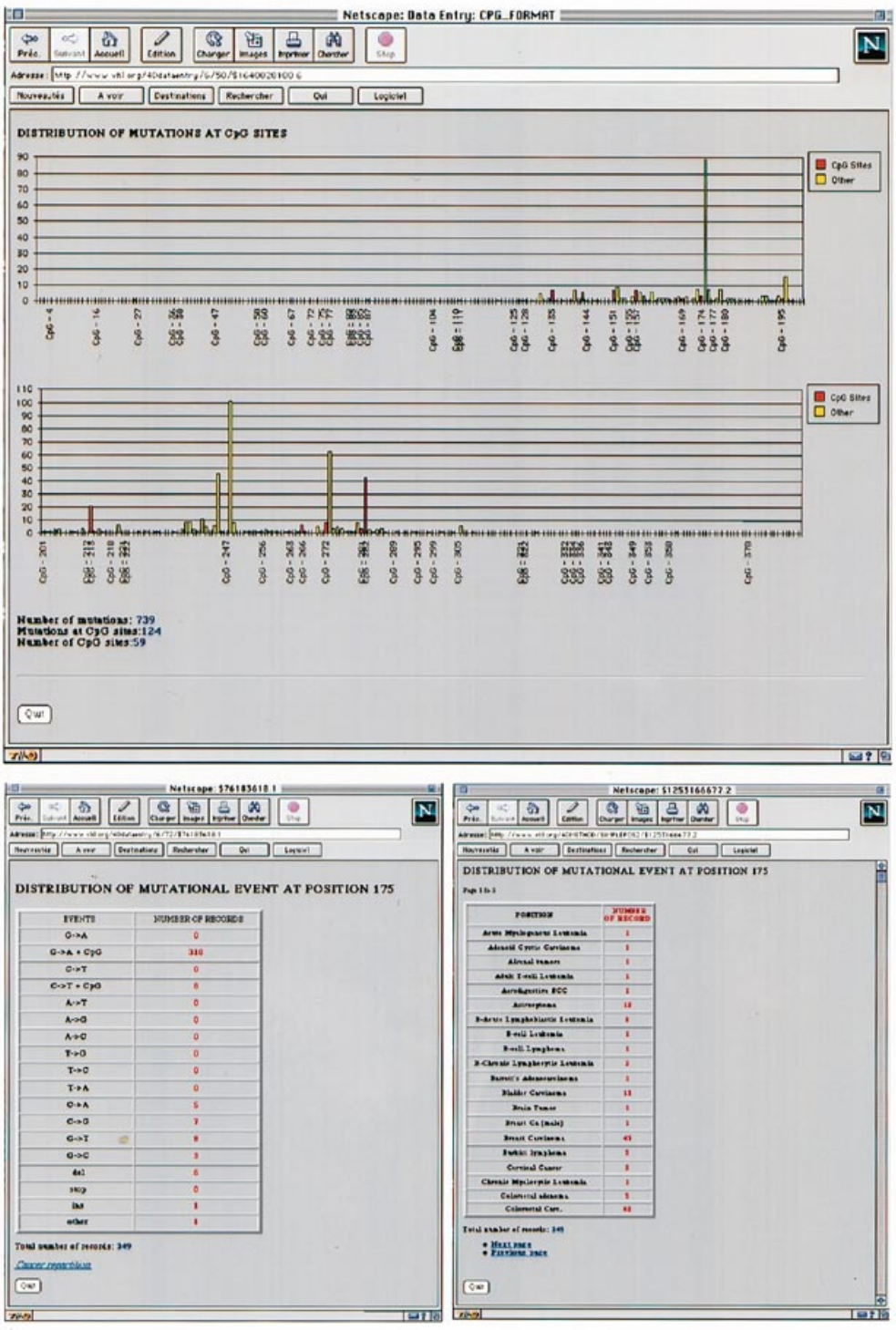

Figure 5. p53 analysis performed directly on the web using the UMD software. In the first example (top of the page) the distribution of p53 mutations is shown along the p53 protein with a special emphasis for the various CpG sites. In the second example (bottom of the page), the number of mutational events and the distribution in various cancers are displayed for a given position. This feature, called 'New mutation?' should help those who want to check their new data.