p53 gene mutation: software and database

Christophe Béroud, Frédérique Verdier[1] and Thierry Soussi*

Hôpital Necker Enfants Malades, U383 INSERM, Paris, France and Université P. et M. Curie, [1]U301 INSERM, 27 rue J. Dodu, 75010 Paris, France

Received September 22, 1995; Revised and Accepted October 23, 1995

ABSTRACT

A large number of different mutations in the tumor suppressor gene p53 gene have been identified in all types of cancer. As of September 1995, this database contains over 4200 mutations. This substantial increase since our previous report can enable epidemiological analyses which were not previously possible. In order to capture all these new data, the software permitting analysis has been improved. This report describes the various improvements since first release of the database.

INTRODUCTION

Over the past few years, progress has been made in cloning genes involved in both monogenic and polygenic disorders, including complex diseases such as cancer (1). Furthermore, for each of these genes, numerous and varied types of alterations have been described, ranging from point mutations to large deletions. A record of the mutations in these various genes serves several important purposes. First, it is clear from all studies performed thus far that mutations are not equally distributed throughout the molecules. Hot spot regions exist which correspond to either a DNA region highly susceptible to mutations (such as CpG dinucleotide), a codon encoding a key residue in the biological function of the protein, or both. Defining such hot spot regions and natural mutants is of invaluable help in defining critical regions in an unknown protein. This information is also valuable for the design of strategy for finding such mutations. In large genes, such as NF1 (59 exons, 2818 amino acids), Rb (27 exons, 928 amino acids), APC (15 exons, 2843 amino acids) and BRCA1 (24 exons, 1863 amino acids), detection of point mutations by direct sequencing analysis is rather difficult due to the size of the target gene. The knowledge of hot spot regions enables focusing on this region, keeping in mind that a negative result should be viewed with caution.

Secondly, it has recently been demonstrated that an alteration in a single gene could cause various types of disorders. Mutations in the RET gene have been associated with multiple endocrine neoplasia types 1 and 2, familial medullary thyroid carcinoma and a non-cancerous disorder known as Hirschprung's disease (2). For each of these disorders, mutations appear to be localized in specific domains of the protein. Furthermore, the location of specific alterations at various positions in a given gene has been shown to be associated with specific clinical features of colon cancer associated with a mutation in the APC gene. A mutation in the C-terminus of the protein has been specifically associated with a secondary abnormality, congenital hypertrophy of the retinal pigment epithelium (3), whereas mutations in the N-terminus are associated with an attenuated phenotype (4).

Thirdly, the design of a database with a large number of point mutations has led to the development of a new field, i.e. molecular epidemiology, where analysis of the mutational spectrum of the various mutations reveals a direct causal effect between carcinogen exposure and a specific cancer (5).

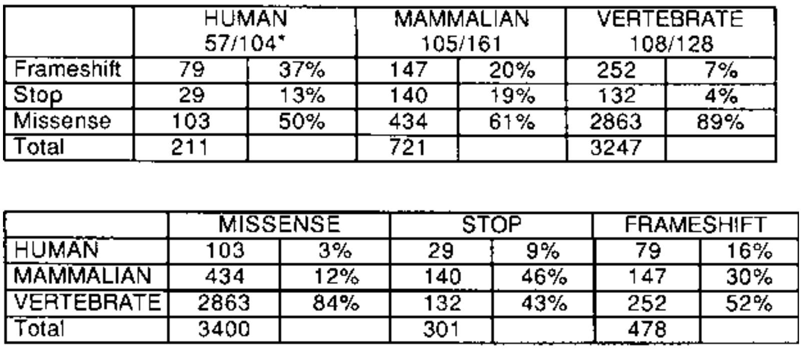
In order to handle this new flow of information, we have devised generic software which allows the entry and analysis of mutations in any gene of interest. It was developed with the 4th Dimension (4D) package from ACI. Although it currently runs on Macintosh, a new version running with Windows will soon be available. This software has been successfully adapted for three genes: p53, APC and fibrillin. A complete description of this software will be given elsewhere (Beroud and Soussi, manuscript in preparation). The present article will describe the latest version of the p53 database while two accompanying papers will specifically describe the APC and fibrillin databases.

VARIATIONS OR MUTATIONS?

As of September 1995, the p53 database contained over 4200 records. Although nonsense and frameshift mutations can easily be considered deleterious for p53 activity, it is more difficult to assess all missense mutations. Among the 393 codons of the human p53 gene, 222 are the targets of 698 different events (excluding nonsense or frameshift mutation). Some codons, such as codon 190, are the target of only one mutational event (C→T, 6 cases) whereas other codons may be the targets of multiple and varies mutational events (codon 179, 69 cases, 9 different events). In most cases, the biological activity of these mutants has not been tested, except for the hot spot codons 175, 248 and 273 (6). Nevertheless, several clues suggest that most of these variations are true mutations which destroy p53 activity. The first indication is the frequency at which a defined codon is mutated, since a high frequency of mutation supports a 'truce' in the alteration. The second clue, from phylogenetic studies, is the fact that the p53 gene has been shown to be highly conserved in vertebrate species (7). Three classes of codon have been distinguished: those specific for human p53, those found in all mammalian p53 and codons conserved in all vertebrates. Structural studies have

* To whom correspondence should be addressed

Table 1. Distribution of p53 mutation according to conservation of the amino acid residue

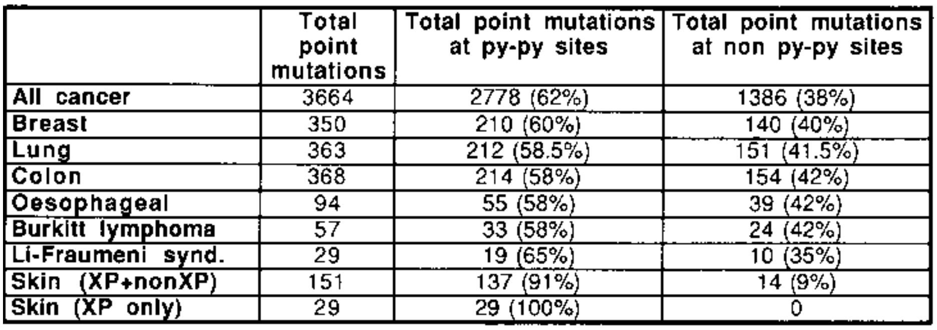


	HUMAN 57/104*		MAMMALIAN 105/161		VERTEBRATE 108/128	
Frameshift	79	37%	147	20%	252	7%
Stop	29	13%	140	19%	132	4%
Missense	103	50%	434	61%	2863	89%
Total	211		721		3247	

	MISSENSE		STOP		FRAMESHIFT	
HUMAN	103	3%	29	9%	79	16%
MAMMALIAN	434	12%	140	46%	147	30%
VERTEBRATE	2863	84%	132	43%	252	52%
Total	3400		301		478	

*Number of residues targeted by mutation. 104 residues are specific for human p53, 161 for mammalian p53 and 128 for vertebrate p53.

Table 2. Distribution of missense and nonsense mutations according to the position of py-py sites



	Total point mutations	Total point mutations at py-py sites	Total point mutations at non py-py sites
All cancer	3664	2778 (62%)	1386 (38%)
Breast	350	210 (60%)	140 (40%)
Lung	363	212 (58.5%)	151 (41.5%)
Colon	368	214 (58%)	154 (42%)
Oesophageal	94	55 (58%)	39 (42%)
Burkitt lymphoma	57	33 (58%)	24 (42%)
Li-Fraumeni synd.	29	19 (65%)	10 (35%)
Skin (XP+nonXP)	151	137 (91%)	14 (9%)
Skin (XP only)	29	29 (100%)	0

confirmed that these conserved amino acids are essential for the DNA binding activity of the protein (8). Table 1 describes the distribution of the various types of mutations among these three classes of codons. This table corresponds to a new feature implemented in the software in order to check the phylogenetic pattern of the mutations. The main finding of this analysis is that 84% of missense mutation target codons are evolutionarily conserved in all p53. Among the remaining 3% (103 cases), which target residues specific for human p53, 11 are found only once at their position, 8 are found twice and the remainder are found more than twice. Similarly, in codons conserved in mammals, only 25 are found only once at their position. All these analyses indicate that essential codons are the true target for p53 inactivation. This database thus provides us with a tremendous amount of information concerning the structure/function relationship of the p53 protein.

UPDATE OF THE SOFTWARE AND THE DATABASE

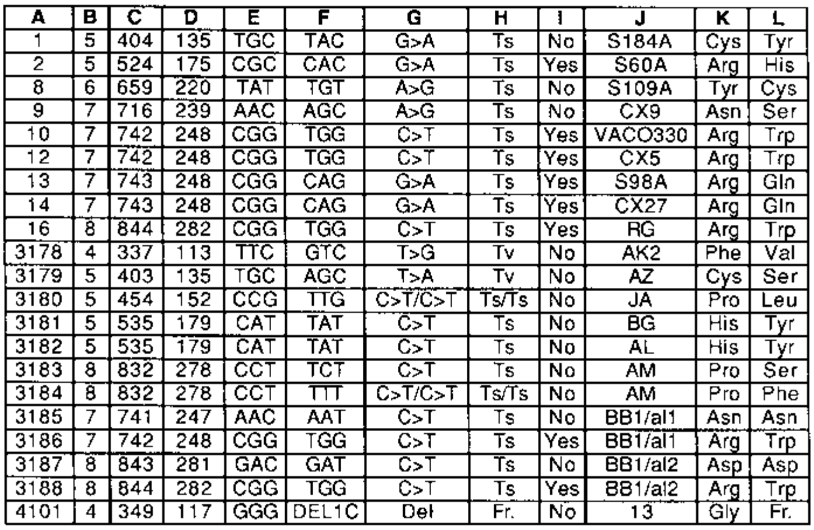
The huge increase in p53 mutations since our last report enables a more precise analysis. For example, the large number of mutations in lung cancer, either SCLC or NSCLC, can be useful in statistical studies of various histological subgroups. In breast cancer, epidemiological studies have led to the discovery that the pattern of the p53 mutation can vary in different geographical areas (9). In skin cancer, analyses of the mutational event in normal individuals or in patients with Xeroderma pigmentosum clearly demonstrate the role of UV as a causal carcinogenic agent in this cancer (5,10) (see Table 2).

In order to take advantage of the increasing possibilities of analysis linked to p53 mutations, new features have been added to the software or to the export format which is provided upon request. Table 2 describes one of the analyses performed which can ascertain the influence of UV in skin cancer. A routine has been developed to distinguish pyrimidine dimers in either the coding or non-coding strand. As stated above, phylogenetic analyses of the various codons have also been added. Due to the increasing number of mutations for several types of cancer, data concerning the histological type have been defined for lung cancer (NSCLC) and breast cancer. All the new features implemented in the software will be described elsewhere (Beroud and Soussi, manuscript in preparation). For each output (table or graphics), an export routine has been set up to recover the data using common software such as Microsoft Word and Excel. The database can be obtained on floppy discs (2 formatted floppy discs are necessary) written in Microsoft Excel format either on an IBM or a Macintosh. Table 3 describes a section of the database in Excel spreadsheet format.

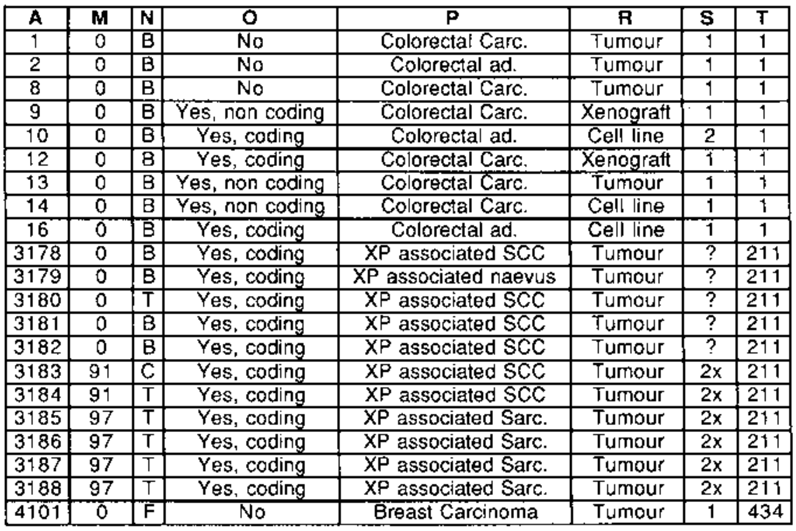
CONCLUSION

Although, the mutation in the p53 gene is the most common alteration in human cancer, it also corresponds to one of the most frequent mutations available for a single gene. From the first compilation, which contained only 350 mutations up to the present one with more than 4200 entries, several new types of analysis are now possible. Such analyses are essential for our

Table 3. Sample listing for p53 mutation database: see footnote for explanation of the various column



A	B	C	D	E	F	G	H	I	J	K	L
1	5	404	135	TGC	TAC	G>A	Ts	No	S184A	Cys	Tyr
2	5	524	175	CGC	CAC	G>A	Ts	Yes	S60A	Arg	His
8	6	659	220	TAT	TGT	A>G	Ts	No	S109A	Tyr	Cys
9	7	716	239	AAC	AGC	A>G	Ts	No	CX9	Asn	Ser
10	7	742	248	CGG	TGG	C>T	Ts	Yes	VACO330	Arg	Trp
12	7	742	248	CGG	TGG	C>T	Ts	Yes	CX5	Arg	Trp
13	7	743	248	CGG	CAG	G>A	Ts	Yes	S98A	Arg	Gln
14	7	743	248	CGG	CAG	G>A	Ts	Yes	CX27	Arg	Gln
16	8	844	282	CGG	TGG	C>T	Ts	Yes	RG	Arg	Trp
3178	4	337	113	TTC	GTC	T>G	Tv	No	AK2	Phe	Val
3179	5	403	135	TGC	AGC	T>A	Tv	No	AZ	Cys	Ser
3180	5	454	152	CCG	TTG	C>T/C>T	Ts/Ts	No	JA	Pro	Leu
3181	5	535	179	CAT	TAT	C>T	Ts	No	BG	His	Tyr
3182	5	535	179	CAT	TAT	C>T	Ts	No	AL	His	Tyr
3183	8	832	278	CCT	TCT	C>T	Ts	No	AM	Pro	Ser
3184	8	832	278	CCT	TTT	C>T/C>T	Ts/Ts	No	AM	Pro	Phe
3185	7	741	247	AAC	AAT	C>T	Ts	No	BB1/al1	Asn	Asn
3186	7	742	248	CGG	TGG	C>T	Ts	Yes	BB1/al1	Arg	Trp
3187	8	843	281	GAC	GAT	C>T	Ts	No	BB1/al2	Asp	Asp
3188	8	844	282	CGG	TGG	C>T	Ts	Yes	BB1/al2	Arg	Trp
4101	4	349	117	GGG	DEL1C	Del	Fr.	No	13	Gly	Fr.



A	M	N	O	P	R	S	T
1	0	B	No	Colorectal Carc.	Tumour	1	1
2	0	B	No	Colorectal ad.	Tumour	1	1
8	0	B	No	Colorectal Carc.	Tumour	1	1
9	0	B	Yes, non coding	Colorectal Carc.	Xenograft	1	1
10	0	B	Yes, coding	Colorectal ad.	Cell line	2	1
12	0	B	Yes, coding	Colorectal Carc.	Xenograft	1	1
13	0	B	Yes, non coding	Colorectal Carc.	Tumour	1	1
14	0	B	Yes, non coding	Colorectal Carc.	Cell line	1	1
16	0	B	Yes, coding	Colorectal ad.	Cell line	1	1
3178	0	B	Yes, coding	XP associated SCC	Tumour	?	211
3179	0	B	Yes, coding	XP associated naevus	Tumour	?	211
3180	0	T	Yes, coding	XP associated SCC	Tumour	?	211
3181	0	B	Yes, coding	XP associated SCC	Tumour	?	211
3182	0	B	Yes, coding	XP associated SCC	Tumour	?	211
3183	91	C	Yes, coding	XP associated SCC	Tumour	2x	211
3184	91	T	Yes, coding	XP associated SCC	Tumour	2x	211
3185	97	T	Yes, coding	XP associated Sarc.	Tumour	2x	211
3186	97	T	Yes, coding	XP associated Sarc.	Tumour	2x	211
3187	97	T	Yes, coding	XP associated Sarc.	Tumour	2x	211
3188	97	T	Yes, coding	XP associated Sarc.	Tumour	2x	211
4101	0	F	No	Breast Carcinoma	Tumour	1	434

Column **A:** Unique mutation identity.
Column **B:** Exon number.
Column **C:** Nucleotide position of the mutation (nucleotide n°1 corresponds to the A residue of the start ATG).
Column **D:** Codon number at which the mutation is located (1–393), numbered as above. If the mutation spans more than one codon, e.g. there is a deletion of several bases, only the first (5′) codon is entered.
Column **E:** Normal base sequence of the codon in which the mutation occurred.
Column **F:** Mutated base sequence of the codon in which the mutation occurred. If the mutation is a base pair deletion or insertion, this is indicated by 'del' or 'ins' followed by the number of bases deleted or inserted and the position of this deletion or insertion in the codon (a, b or c). The nucleotide position is the first that is deleted or the one following the insertions. For example, 'del66b' is a deletion of 66 bases including the second base of the codon; 'ins4b' is an insertion of 4 bases occurring between the first and second base of the codon.
Column **G:** Give the base change, read from the coding strand by convention, for base substitutions.
Column **H:** Mutational event (transition/transversion or frameshift).
Column **I:** Indicate whether the mutation is a transition occurring at a CpG dinucleotide.
Column **J:** Name of the tumor/patient/cell line as given by the authors.
Column **K:** Wild type amino acid.
Column **L:** Mutant amino acid. Deletion and insertion mutations which results in frameshift are designated by 'Fr'.
Column **M:** Special identification of mutations which are found more than once in a given tumor.
Column **N:** mutation type: B, single mutation; D, deletion 1 bp; I, insertion 1 bp; F, del or ins ≥2; T, tandem mutation; C, complex mutation.
Column **O:** Localization of the mutation with respect to a pyrimidine dimer (yes or no) in either the coding or the non coding strand.
Column **P:** Cancer.
Column **Q:** Stage [for several cancers such as lung, breast or colorectal carcinomas, the stage is included (if available)].
Column **R:** Origin of the mutation (tumor, cell line, xenograft or germline).
Column **S:** LOH, if available. 2, two alleles are remaining; 1, only one allele is remaining; ?, no information available or non informative; 2x, two mutations in the same tumor at different positions.
Column **T:** Reference number.

understanding of the effect of the environment, carcinogen exposure and their relationship to cancer development.

ACKNOWLEDGEMENTS

We are grateful to J. Bram for reading the manuscript. This work was supported by grants from the Association de Recherche sur le Cancer, Ligue Nationale contre le Cancer (Comité de Paris) and Ligue Nationale contre le Cancer (Comité National).

REFERENCES

1 Collins, F.S. (1995) *Nature Genet.* **9**, 347–350.

2 Smith, D.P., C. Eng and B.A.J. Ponder (1994) *J. Cell Sci.* 43–49.

3 Olschwang, S., A. Tiret, P. Laurentpuig, M. Muleris, R. Parc and G. Thomas (1993) *Cell* **75**, 959–968.

4 Spirio, L., S. Olschwang, J. Groden, M. Robertson, W. Samowitz, G. Joslyn, L. Gelbert, A. Thliveris, M. Carlson, B. Otterud, H. Lynch, P. Watson, P. Lynch, P. Laurentpuig, R. Burt, J.P. Hughes, G. Thomas, M. Leppert and R. White (1993) *Cell* **75**, 951–957.

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

6 Ory, K., Y. Legros, C. Auguin and T. Soussi (1994) *EMBO J.* **13**, 3496–3504.

7 Soussi, T., C. Caron de Fromentel and P. May (1990) *Oncogene* **5**, 945–952.

8 Cho, Y.J., S. Gorina, P.D. Jeffreyand N.P. Pavletich (1994) *Science* **265**, 346–355.

9 Saitoh, S., J. Cunningham, E.M.G. Devries, R.M. McGovern, J.J. Schroeder, A. Hartmann, H. Blaszyk, L.E. Wold, D. Schaid, S.S. Sommer and J.S. Kovach (1994) *Oncogene* **9**, 2869–2875.

10 Ziegler, A., A.S. Jonason, D.J. Leffell, J.A. Simon, H.W. Sharma, J. Kimmelman, L. Remington, T. Jacks and D.E. Brash (1994) *Nature* **372**, 773–776.