APC gene: database of germline and somatic mutations in human tumors and cell lines

Christophe Béroud and Thierry Soussi[1],*

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

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

ABSTRACT

A database is described in which over 700 mutations in the human APC gene of tumors (colon cancer predominantly) are compiled from the literature. It includes both molecular informations about the mutations and also clinical data about the patients. A software have been designed in order to analyse all these informations in the database.

INTRODUCTION

Familial adenomatous polyposis (FAP) is an autosomal-dominant precancerous condition characterized by the appearance of hundreds to thousands of adenomatous polyps throughout the entire colorectum (1). The disease is caused by a germline mutation in the tumor suppressor gene APC localized on chromosome 5q21–22. Furthermore, somatic mutation of the APC gene has been identified in sporadic colorectal cancer as well as in some cancer of the stomach, pancreas, thyroid, ovary and other primary sites (2). As shown below, >98% of APC mutations are either frameshift or nonsense mutations leading to the synthesis of a truncated protein.

Some interesting features which were observed by analysis of the APC mutation concern the possible relationship between the localization of the mutation and specific phenotypes associated with the disease. Nagase *et al.* have shown a striking correlation between the location of the mutation and the number of polyps in FAP patients (3). In profuse types of FAP patients, >5000 adenomatous polyps and mutations (frameshift mainly) are concentrated between codon 1255 and 1467. In sparse types of FAP patients, the number of polyps is comprised between 1000 and 2000, and APC mutations, missense and frameshift mutations, are localized in other regions of the APC protein, mainly the 5′ half of the protein. A retinal lesion, called congenital hypertrophy of the retinal pigment epithelium (CHRPE) is frequently associated with FAP. Recently, it has been reported that the severity of CHRPE is dependent on the position of the mutation (4). Retinal lesions were usually absent when mutations occurred from exon 1 to exon 9, and present when they occurred 3′ to it. Finally, it seems that the presence of germline mutations in the extreme 5′ of the APC coding region (before codon 157) are correlated with an attenuated phenotype (fewer polyps and the delayed onset of the cancer) (5). Although all these correlations are striking, more works are needed to strengthen them.

Taken together, these observations strongly suggest a correlation between several phenotypes associated with the disease and the location of the mutation. Therefore, we established a specific database devoted to APC mutation. The software package contains routines for analysis of the APC database, developed using the 4th dimension″ (4D) package from ACI. Use of the 4D SGDB gives access to optimized multicriteria research and sorting tools to select records from any field. Moreover, several routines were specifically developed: (i) the routine entitled ‘position’ examines the distribution of mutations at the nucleotide level to identify preferential mutation sites; (ii) ‘statistical evaluation of mutational events’ is comparable to ‘position’, but also indicates the type of mutational event. The result can either be displayed in tables or in a graphic representation (see Table 1); (iii) ‘frequency of mutations’ enables the study of the relative distribution of mutations at all sites, and sorts them according to their frequency. A graphic representation is available and displays a cumulative chart of distribution of mutations (Fig. 1); (iv) ‘stat exons’ studies the distribution of mutations in the different exons and enables detection of a statistically significant differences between observed and expected mutations. The software also includes the analysis of correlations between clinical data and localization of the mutation.

MAJOR FINDINGS FOR APC MUTATION ANALYSIS

More than 700 mutations have been reported thus far, and several interesting features are revealed by their analysis. As described in Table 1, most of the mutations lead to truncation of the APC protein either by a nonsense mutation (30%) or by a frameshift mutation (68%). This feature clearly indicates that each mutation is truly deleterious for APC function. Most of the mutations (germline or somatic) occurred in the first half of the coding region. Germline mutations are scattered throughout the 5′ half of the gene, whereas the somatic mutations (60%) are concentrated in a region called the mutation cluster region (MCR). In germline mutation, two hot spot codons have been identified, one at position 1061 and the second at position 1309. In somatic mutations, two hot spots seem to occur at position 1309 and 1450 (Fig. 1). This concentration of mutations in the 5′ part of the gene, leading to the synthesis of truncated APC, is believed to be

[*] To whom correspondence should be addressed

Table 1. Summary of germline and somatic mutations of the APC gene in FAP and non FAP patients

	TOTAL	Germline	Somatic
Total	737	332	402
Frameshifts			
deletion	432 (58%)	220 (66%)	210 (52%)
insertion	73 (10%)	21 (6%)	52 (13%)
point mutation			
missense	16 (2%)	6 (2%)	12 (3%)
nonsense	216 (30%)	85 (26%)	129 (32%)
G->A	7	5	2
G->A at CpG	1	0	1
C->T	44	17	26
C->T at CpG	94	38	56
A->T	8	5	3
A->G	5	0	5
A->C	0	0	0
T->G	5	2	3
T->C	2	0	2
T->A	3	2	1
C->A	19	9	10
C->G	15	10	5
G->T	28	3	24
G->C	1	0	1

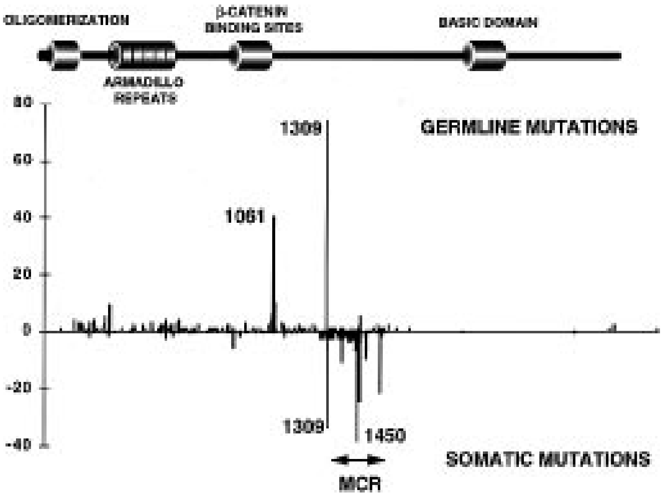


Figure 1. Distribution of germline and somatic mutations in the APC gene. (Top) Linear diagram showing the relative positions of the various important feature of the APC protein.

involved in a dominant effect of the N-terminus of the APC, protein throughout the entire protein (6). This APC region contains a dimerization domain and it has been demonstrated that wild type and mutant APC are associated *in vivo,* thereby demonstrating a potential for dominant interference by the truncated mutant protein. Although the precise biological function of APC is not known, it has been shown that APC protein associates with catenins or with microtubules. Domains of the APC protein involved in these interactions have been mapped in the second half of the APC protein and are therefore missing in the mutant APC (7 for review).

Another unusual finding came about from the analysis of transition at the CpG dinucleotide. This dinucleotide is known to be frequently methylated in the vertebrate genome. Deamination of the 5-methylcytosine can generate a C→T transition after DNA replication. The symmetry of dinucleotide CpG suggests that the 5-methylcytosine residue on each strand of the DNA can be the target of deamination. If we consider this mutational event on a given strand of DNA (usually the coding strand), mutational events can be depicted as a C→T or G→A transition occurring at an equal rate. In the p53 gene, three amino acids (Arg175, Arg248 and Arg273) account for 22% of all mutations. These three codons (CGN) contain a CpG dinucleotide. For codons 248 and 273, two types of mutational event are equally represented, e.g. 69 G→A and 79 C→T for codon 248 and 73 G→A and 63 C→T for codon 273. In contrast, there is a striking imbalance for codon 175, with 92 G→A (Arg→His) and only 2 C→T (Arg→Cys). This low incidence of Cys175 mutant has been explained by the observations that the biological function of p53 is not inactivated by this mutation (8). Examination of the APC missense mutations at CpG dinucleotides reveals a striking bias: 99% (94/95) are C→T transitions, whereas only 1% (1/95) is a G→A. This bias can be explained by the fact that APC protein has to be truncated in order to be inactivated. Most of these transitions (92/93) occur at CGA codons, as the transition leads to the nonsense mutation TGA. This observation highlights the selection pressure on the APC gene for making a truncated protein in order to play a role in colon cancer. However, it also means that despite the high number of APC gene mutations available, such information is not suitable for molecular epidemiology, as was the case for the p53 gene.

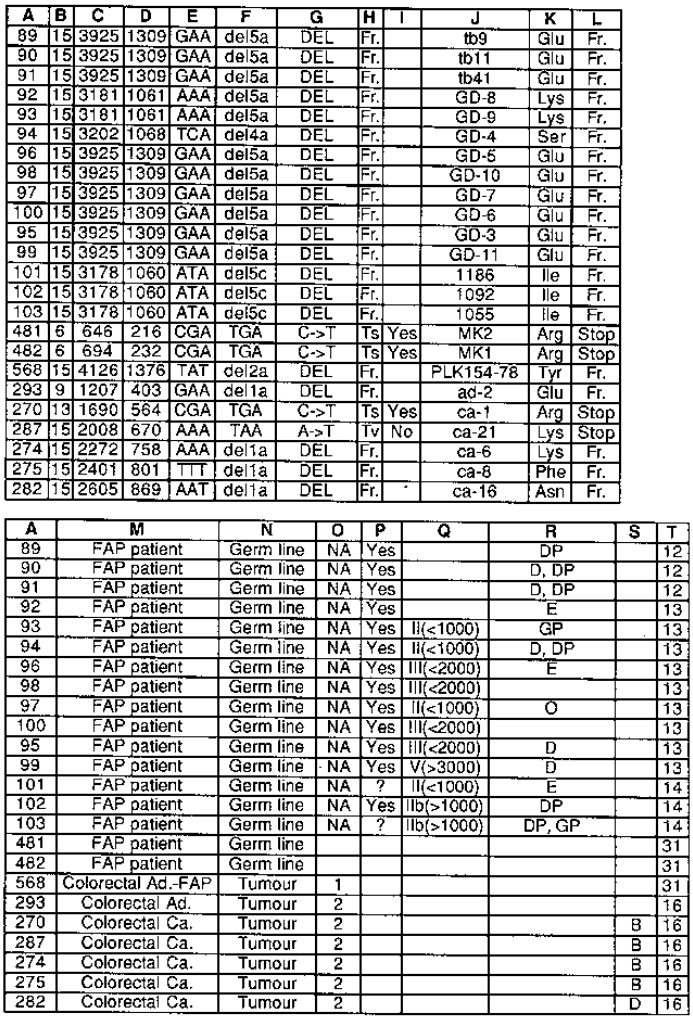
DESCRIPTION OF THE DATABASE AND FORMAT

As of August 1995, the data base contained 738 records of mutations, either germline, somatic or from cell lines. Splice mutations were omitted. When the same mutation was reported in more than one article, only the first report was taken into account. Most of the mutations described in this database originated either from FAP patients or from colorectal cancers. For germline mutations in FAP patients, each record corresponded to only one patient in a given family. Relatives in the family for whom the same mutation had been observed were not recorded. For somatic mutations in colorectal cancer, several authors reported that different mutations could occur in different carcinomas in a single patient (either in dysplasias, adenomas or carcinomas). For each mutation, a single record was entered, corresponding to a single mutational event. More recently, somatic mutations in the APC gene have been described in other neoplasms associated either with the gastric tract (pancreas, stomach) or other sites such as thyroid or ovary. Such mutations were also included in the database, but their number is too low for statistical analysis. More than 20 polymorphisms have been identified in the APC gene, and have been compiled by Nagase *et al.* (9). They have not been taken into account in the database. Table 2 describes a section of the database in Excel spreadsheet format.

AVAILABILITY

The database can be obtained on floppy discs (two formatted floppy discs are necessary) written in Microsoft Excel either on

Table 2. Sample listing for APC mutation database: see footnote for explanation of the various column



A	B	C	D	E	F	G	H	I	J	K	L
89	15	3925	1309	GAA	del5a	DEL	Fr.		tb9	Glu	Fr.
90	15	3925	1309	GAA	del5a	DEL	Fr.		tb11	Glu	Fr.
91	15	3925	1309	GAA	del5a	DEL	Fr.		tb41	Glu	Fr.
92	15	3181	1061	AAA	del5a	DEL	Fr.		GD-8	Lys	Fr.
93	15	3181	1061	AAA	del5a	DEL	Fr.		GD-9	Lys	Fr.
94	15	3202	1068	TCA	del4a	DEL	Fr.		GD-4	Ser	Fr.
96	15	3925	1309	GAA	del5a	DEL	Fr.		GD-5	Glu	Fr.
98	15	3925	1309	GAA	del5a	DEL	Fr.		GD-10	Glu	Fr.
97	15	3925	1309	GAA	del5a	DEL	Fr.		GD-7	Glu	Fr.
100	15	3925	1309	GAA	del5a	DEL	Fr.		GD-6	Glu	Fr.
95	15	3925	1309	GAA	del5a	DEL	Fr.		GD-3	Glu	Fr.
99	15	3925	1309	GAA	del5a	DEL	Fr.		GD-11	Glu	Fr.
101	15	3178	1060	ATA	del5c	DEL	Fr.		1186	Ile	Fr.
102	15	3178	1060	ATA	del5c	DEL	Fr.		1092	Ile	Fr.
103	15	3178	1060	ATA	del5c	DEL	Fr.		1055	Ile	Fr.
481	6	646	216	CGA	TGA	C->T	Ts	Yes	MK2	Arg	Stop
482	6	694	232	CGA	TGA	C->T	Ts	Yes	MK1	Arg	Stop
568	15	4126	1376	TAT	del2a	DEL	Fr.		PLK154-78	Tyr	Fr.
293	9	1207	403	GAA	del1a	DEL	Fr.		ad-2	Glu	Fr.
270	13	1690	564	CGA	TGA	C->T	Ts	Yes	ca-1	Arg	Stop
287	15	2008	670	AAA	TAA	A->T	Tv	No	ca-21	Lys	Stop
274	15	2272	758	AAA	del1a	DEL	Fr.		ca-6	Lys	Fr.
275	15	2401	801	TTT	del1a	DEL	Fr.		ca-8	Phe	Fr.
282	15	2605	869	AAT	del1a	DEL	Fr.		ca-16	Asn	Fr.

A	M	N	O	P	Q	R	S	T
89	FAP patient	Germ line	NA	Yes		DP		12
90	FAP patient	Germ line	NA	Yes		D, DP		12
91	FAP patient	Germ line	NA	Yes		D, DP		12
92	FAP patient	Germ line	NA	Yes		E		13
93	FAP patient	Germ line	NA	Yes	II(<1000)	GP		13
94	FAP patient	Germ line	NA	Yes	II(<1000)	D, DP		13
96	FAP patient	Germ line	NA	Yes	III(<2000)	E		13
98	FAP patient	Germ line	NA	Yes	III(<2000)			13
97	FAP patient	Germ line	NA	Yes	II(<1000)	O		13
100	FAP patient	Germ line	NA	Yes	III(<2000)			13
95	FAP patient	Germ line	NA	Yes	III(<2000)	D		13
99	FAP patient	Germ line	NA	Yes	V(>3000)	D		13
101	FAP patient	Germ line	NA	?	II(<1000)	E		14
102	FAP patient	Germ line	NA	Yes	IIb(>1000)	DP		14
103	FAP patient	Germ line	NA	?	IIb(>1000)	DP, GP		14
481	FAP patient	Germ line						31
482	FAP patient	Germ line						31
568	Colorectal Ad.-FAP	Tumour	1					31
293	Colorectal Ad.	Tumour	2					16
270	Colorectal Ca.	Tumour	2				B	16
287	Colorectal Ca.	Tumour	2				B	16
274	Colorectal Ca.	Tumour	2				B	16
275	Colorectal Ca.	Tumour	2				B	16
282	Colorectal Ca.	Tumour	2				D	16

Column **A:** Unique mutation identity.
Column **B:** Exon number.
Column **C:** Nucleotide position of the mutation (nucleotide n°l correspond 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 span 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 the 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/transvertion or Frameshift).
Column **I:** Indicate if 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 result in frameshift are designated by 'Fr'.
Column **M:** Cancer.
Column **N:** Origin of the mutation (tumor, cell line, xenograft or germline).
Column **O:** LOH, if available. 2, two alleles are remaining; 1, only one allele is remaining; ?, no information available or non informative.
Columns P to S correspond to clinical information if available:
Column **P:** CHRPE for FAP patients.
Column **Q:** Number of polyps for FAP patients.
Column **R:** Extracolonic lesions: O, osteoma; E, epidermoid cyst; D, desmoid; T, thyroid tumor; GP, gastric polyp; DP; duodenal polyp; DC, duodenal cancer.
Column **S:** Stage of the tumor according to Dukes staging (A to D).
Column **T:** Reference number.
NA: Not applicable.

an IBM or a Macintosh. Notification of omissions and errors in the current version will be gratefully appreciated by the corresponding author.

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 the Ligue Nationale contre le Cancer (Comité National).

REFERENCES

1 Utsunomiya, J. and Lynch, H.L. (1990) *Hereditary Colorectal Cancer.* Springer Verlag, Tokyo, pp. 3–16.
2 Nakamura, Y. (1993) *Adv. Cancer Res.* **62**, 65–87.
3 Nagase, H., Y. Miyoshi, A. Horii, T. Aoki, M. Ogawa, J. Utsunomaya, S. Baba, T. Sasazuki and Y. Nakamura (1992) *Cancer Res.* **52**, 4055–4057.
4 Olschwang, S., A. Tiret, P. Laurent-Puig, M. Muleris, R. Parc and G. Thomas (1993) *Cell* **75**, 959–968.
5 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. Laurent-Puig, R. Burt, J.P. Hughes, G. Thomas, M. Leppert and R. White (1993) *Cell* **75**, 951–957.
6 Su, L.K., K.A. Johnson, K.J. Smith, D.E. Hill, B. Vogelstein and K.W. Kinzler (1993) *Cancer Res.* **53**, 2728–2731.
7 Polakis, P. (1995) *Curr. Opin. Gen. Dev.* **5**, 66–71.
8 Ory, K., Y. Legros, C. Auguin and T. Soussi (1994) *EMBO J.* **13**, 3496–3504.
9 Nagase, H. and Y. Nakamura (1993) *Hum. Mutat.* **2**, 425–34.