HUMAN MUTATION 15:86–94 (2000)

MDI SPECIAL ARTICLE

UMD (Universal Mutation Database): A Generic Software to Build and Analyze Locus-Specific Databases

Christophe Béroud,[1]* Gwenaëlle Collod-Béroud,[1] Catherine Boileau,[1,2] Thierry Soussi,[3] and Claudine Junien[1,2]

[1]*INSERM U383, Hôpital Necker Enfants Malades, Clinique Maurice Lamy, Paris, France*
[2]*Laboratoire de Biochimie, d'Hormonologie et de Génétique Moléculaire, Hôpital Ambroise Paré, Boulogne, France*
[3]*Institut Curie, UMR 218 CNRS, Pavillon Trouillet Rossignol, Paris, France*

The human genome is thought to contain about 80,000 genes and presently only 3,000 are known to be implicated in genetic diseases. In the near future, the entire sequence of the human genome will be available and the development of new methods for point mutation detection will lead to a huge increase in the identification of genes and their mutations associated with genetic diseases as well as cancers, which is growing in frequency in industrial states. The collection of these mutations will be critical for researchers and clinicians to establish genotype/phenotype correlations. Other fields such as molecular epidemiology will also be developed using these new data. Consequently, the future lies not in simple repositories of locus-specific mutations but in dynamic databases linked to various computerized tools for their analysis and that can be directly queried on-line. To meet this goal, we devised a generic software called UMD (Universal Mutation Database). It was developed as a generic software to create locus-specific databases (LSDBs) with the 4th Dimension® package from ACI. This software includes an optimized structure to assist and secure data entry and to allow the input of various clinical data. Thanks to the flexible structure of the UMD software, it has been successfully adapted to nine genes either involved in cancer (APC, P53, RB1, MEN1, SUR1, VHL, and WT1) or in genetic diseases (FBN1 and LDLR). Four new LSDBs are under construction (VLCAD, MCAD, KIR6, and COL4A5). Finally, the data can be transferred to core databases. Hum Mutat 15:86–94, 2000.

KEY WORDS: **MDI; UMD; mutation database; software; data analysis**

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. Indeed, the recent increase of sequence data, a consequence of large sequencing projects like the Human Genome Project, and the enormous growth in data from expressed sequence tags [Benson et al., 1997] have become essential tools to localize disease genes. For each disease gene, numerous and varied types of alterations have been described, ranging from point mutations to large deletions. With the development of our knowledge in gene alterations leading to human diseases, it has become clear that the identification of these mutations should play a critical role not only in diagnosis and prognosis but also in research.

First, it is clear from all studies performed so far that mutations are generally not randomly distributed. Hot spot regions exist which correspond to either a DNA region highly susceptible to mutations (such as CpG dinucleotides), or 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. In large genes, such as NF1 (59 exons, 2,818 amino acids, MIM# 162200), Rb (RB1; 27 exons, 928 amino acids, MIM# 180200), APC (15 exons, 2,843 amino acids, MIM# 175100), and BRCA1 (24 exons, 1,863 amino acids, MIM# 113705), detection of point mutations by direct sequencing analy-

Received 27 July 1999; accepted revised manuscript 8 October 1999.

*Correspondence to: Christophe Béroud, INSERM U383, Hôpital Necker-Enfants Malades, Pavillon Maurice Lamy, 149-161 rue de Sèvres, 75743 Paris Cedex 15, France. E-mail: beroud@necker.fr

sis is rather difficult due to the size of the target gene. The knowledge of a hot spot region enables focusing on this region, keeping in mind that a negative result should be viewed with caution.

Second, it is now well demonstrated that alterations in a single gene can cause various types of disorders, like the RET gene (MIM# 164761). Mutations in the latter have been associated with multiple endocrine neoplasia types IIA [Mulligan et al., 1993] and IIB [Hofstra et al., 1994], familial medullary thyroid carcinoma [Xue et al., 1994] and a noncancerous disorder known as Hirshprung's disease [Edery et al., 1994; Romeo et al., 1994]. 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, as is the case of colon cancer and mutations 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 [Olschwang et al., 1993], whereas mutations in the N-terminus are associated with an attenuated phenotype [Spirio et al., 1993].

Third, analysis of mutations can lead to the definition of risk factors. For instance, in the VHL (MIM# 193300) families presenting mutations leading to truncated proteins, at least one member developed an RCC in 83% of cases vs. only 54% in families presenting missense mutations [Gallou, 1999].

Fourth, in diseases characterized by considerable variation in the clinical phenotype between families and also within the same family, like Marfan syndrome (FBN1; MIM# 154700), it is of great importance to confirm or firmly exclude the diagnosis in at risk family members as early as possible because of the potential fatal cardiovascular complications of the disease.

Finally, the analysis of the p53 database (TP53), which contains a very large number of point mutations, has led to the development of a new field, i.e., molecular epidemiology, where the analysis of the mutational spectrum reveals a direct causal effect between carcinogen exposure and a specific cancer [see Soussi et al., 2000].

All these examples show that the future lies not in simple repositories of locus-specific mutations but in dynamic databases linked to various computerized tools for their analysis, and that can be directly queried on-line. To meet this goal, we devised a generic software called UMD (Universal Mutation Database).

DATABASE STRUCTURE

The UMD software was developed not only to create various locus-specific mutation databases (LSDBs) but also numerous analyzing tools. It was thus critical to optimize the database structure and use an appropriate programming tool to meet these two goals. Among the various languages available, we chose the 4th Dimension® package from ACI because it had many advantages: 1) it runs on both Macintosh and PC platforms, 2) it allows the creation of relational databases, 3) it has a complete language of more than 700 commands, 4) it has a graphic interface and can create dynamic HTML pages, giving easy access to direct on-line queries via the Web, and 5) it has a compiler to create optimized software.

To avoid the many errors found in publications of mutations (up to 10%) which include wrong nucleotide or AA position, reference to a wrong sequence or misinterpretation of the mutation, we built a specific structure including two tables called "Genetic Code" and "Gene Sequence." The "Genetic Code" table is common to all LSDBs developed with UMD and contains the human "codon usage" genetic code (it can be changed if applications are made to other species). For each codon the amino acid (AA) three letter code and the amino acid mutability value are available. We defined this new parameter that is calculated as follows: for each base, the number of relevant substitutions (leading to AA change) is evaluated (0–3) and values for the three bases of a specific codon are added (for codon CUA the mutability value is 5 while for codon UGA it is 9). The "Gene Sequence" is specific for each LSDB. It includes for each AA position the wild-type codon and phylogenic data, defined by curators using a numeric value (conserved AA among mammals (value = 2), vertebrates (value = 3) ...). The use of these two tables secures the data entry, avoiding typing or numbering errors.

Concurrently, a "Mutation" and a "Clinical Data" table were created. The "Mutation" table is the central part of the structure. It is linked to all other tables. It includes many data, with minor differences from one LSDB to another. To reduce typing errors and facilitate the input of data, the software automatically checks and calculates various data, such as wild-type codon and AA, mutant AA, exon number, mutational event, mutation type, involvement of a CpG or a pyrimidine dou-

blet, localization of the mutation in a highly conserved domain and/or in a structural domain and modification of the restriction map. Finally, when the mutation is a deletion or an insertion, the UMD software searches automatically for the involvement of a repeated sequence that could account for the mutational event. Recently, we added a routine to automatically name the mutation according to the international nomenclature [Antonarakis et al., 1998; den Dunnen and Antonarakis, 2000].

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, as shown in Table 1 [for details see Béroud et al., 1996; Béroud and Soussi, 1997, 1998; Collod et al., 1996; Collod-Béroud et al., 1997, 1998].

LSDBS DEVELOPED WITH THE UMD SOFTWARE

To our knowledge, UMD is the first software that can be used to create gene-specific mutation databases and to analyze data either on a personal com-

TABLE 1. LSDB Routines

Menus-functions	APC	FBN1	LDLR	MEN1	p53	SUR1	VHL	WT1
FILE								
Export data	Yes	—	Yes	Yes	Yes	Yes	Yes	Yes
Modify structure	Yes	—	Yes	Yes	Yes	Yes	Yes	Yes
Update clinical data	Yes	—	Yes	Yes	–	Yes	Yes	Yes
REFERENCE								
Add new reference	Yes	—	Yes	Yes	Yes	Yes	Yes	Yes
Show all references	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Search by authors	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Search by keyword	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Search Medline ID	Yes	—	Yes	Yes	Yes	Yes	Yes	Yes
Search by reference number	Yes	—	Yes	Yes	Yes	Yes	Yes	Yes
MUTATION								
Add new record	Yes	—	Yes	Yes	Yes	Yes	Yes	Yes
AA type search	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Reference search	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Deletion analysis	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Insertion analysis	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SHOW ALL MUTATIONS								
Generic display	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mutagenesis display	Yes	Yes	—	Yes	Yes	Yes	Yes	Yes
Structural display	Yes	Yes	—	Yes	Yes	Yes	Yes	Yes
Clinical display	Yes	Yes	—	Yes	Yes	Yes	Yes	Yes
Free search	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
STATVIEW								
Position	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mutational events	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Detailed mutation events	Yes	—	Yes	Yes	Yes	Yes	Yes	Yes
Frequency of mutations	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Frequency of events	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Distribution of mutations	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Binay comparison	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stat exons	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Distribution by exons	Yes	—	Yes	Yes	Yes	Yes	Yes	Yes
UV induced mutations	Yes	—	Yes	Yes	Yes	Yes	Yes	Yes
Tandem mutations	Yes	—	Yes	Yes	Yes	Yes	Yes	Yes
Coding strand mutations	Yes	—	Yes	Yes	Yes	Yes	Yes	Yes
Phylogeny	Yes	—	Yes	Yes	Yes	Yes	Yes	Yes
Structure	Yes	—	Yes	Yes	Yes	Yes	Yes	Yes
Partition of mutations	Yes	—	Yes	Yes	Yes	Yes	Yes	Yes
Cancer distribution	—	—	—	—	Yes	—	—	—
PROTEIN								
NH2 unique region	—	Yes	—	—	—	—	—	—
EGF-module	—	Yes	—	—	—	—	—	—
Hybrid module	—	Yes	—	—	—	—	—	—
Cb EGF module	—	Yes	—	—	—	—	—	—
8-Cystein module	—	Yes	—	—	—	—	—	—
Proline rich region	—	Yes	—	—	—	—	—	—
COOH unique region	—	Yes	—	—	—	—	—	—
Ligand binding	—	—	Yes	—	—	—	—	—
EGF Precursor like region	—	—	Yes	—	—	—	—	—

puter or via the Internet. Contrary to other databases, this is the first interface that provides the possibility of analyzing data and displaying results in a graphic representation. Because of the flexible structure of the UMD software, it has been successfully adapted for 13 genes: TP53 (11,103 mutations, MIM# 191170) [Béroud et al., 1996; Béroud and Soussi, 1997, 1998], APC (1,612 mutations) [Béroud and Soussi, 1996b, 1997; Laurent-Puig et al., 1998], FBN1 (169 mutations) [Collod et al., 1996; Collod-Béroud et al., 1997, 1998], LDLR (525 mutations, MIM# 143890) [Varret et al., 1997, 1998], VHL (659 mutations) [Béroud et al., 1998], WT1 (148 mutations, MIM# 194070) [Jeanpierre et al., 1998], SUR1 (97 mutations, MIM# 600509) (J.C. Fournet, personal communication), MEN1 (184 mutations, MIM# 131100) (A. Callender, personal communication) and RB1 (303 mutations) (F. Namouni, personal communication). Four new LSDBs are under construction: VLCAD (MIM# 201475) and MCAD (MIM# 201450) (B. Storstein Andresen, Denmark), KIR6 (MIM# 600937) (International Consortium), and COL4A5 (MIM# 120131) (International Consortium).

Eight LSDBs are accessible via Internet and the Word Wide Web interfaces (http://www.umd.necker.fr): APC, FBN1, LDLR, MEN1, TP53, SUR1, VHL, and WT1. The UMD software is freely available and can be downloaded from ftp.umd.necker.fr.

The UMD software was first developed for the p53 gene [see Soussi et al., 2000] and subsequently adapted to many genes involved in cancer. The creation of the first LSDB for a genetic disease (FBN1 database) revealed the need for some modifications, such as the addition of clinical data. Today the UMD software has been applied to two fields of genes/diseases with specific input data and analyzing tools: cancer and genetic disease databases.

CANCER DATABASES

Today six LSDBs have been created for tumor suppressor genes involved in various cancers. All these LSDBs share data and routines (see Tables 1, 2). The p53 database is a model for molecular epidemiology studies and most routines were first developed for this gene and then included in all subsequent cancer gene UMD softwares. Specific

TABLE 2. Fields Available in 8 LSDBs

	APC	FBN1	LDLR	MEN1	p53	SUR1	VHL	WT1
AA position, nucleotide, sequence search	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
WT codon,* WT AA*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mutant codon, mutant AA*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Exon*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mutational event*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Type*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
CpG*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Py-Py doublet*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Structure*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Insertion*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Highly conserved domain*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Polymorphism or mutation*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sample name	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Patient	—	—	—	—	Yes	—	—	—
Origin	Yes	Yes	–	Yes	Yes	Yes	Yes	Yes
Pathology	—	Yes	—	—	—	Yes	—	—
Age of onset	—	Yes	—	—	—	Yes	—	—
Mutation number	—	—	—	—	—	Yes	—	—
Tumor	—	—	—	Yes	—	—	—	—
Cancer	Yes	—	—	—	Yes	—	Yes	Yes
Organ	—	—	—	—	Yes	—	—	—
Histology	Yes	—	—	—	Yes	—	Yes	Yes
Stage	Yes	—	—	—	Yes	—	Yes	Yes
LOH	Yes	—	—	—	Yes	—	Yes	Yes
Mutation type*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Smoking	Yes	—	—	Yes	Yes	—	Yes	Yes
Aflatoxin	—	—	—	—	Yes	—	—	—
Diazoxide status	—	—	—	—	—	Yes	—	—
Skin tumor or Internal tumor	—	—	—	Yes	Yes	—	Yes	Yes
Leukemia/Lymphoma or Solid tumor	—	—	—	Yes	Yes	—	Yes	Yes
Reference	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clinical data	Yes	Yes	Yes	Yes	—	Yes	Yes	Yes
Comments	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

*These fields are automatically filled in by the software.

questions can be addressed for these genes: Is a carcinogenic substance involved in the etiology of a tumor? Is it a specific carcinogen such as aflatoxin B1 or UV-induced mutations? Are some specific mutations associated with a specific stage, grade, or subtype of the tumor? As all questions cannot be addressed in this article, we will limit our example to the analysis of the mutational events in search of a carcinogenic exposure. For this, we use the tool "Partition of Mutations" for two groups of records: germline mutations vs. somatic mutations. Results are shown in Table 3. We observe a difference in the distribution of mutations in these two groups. In fact, we can define three categories of genes: the first includes APC and shows a majority of mutations leading to truncated proteins (MLTP) (including nonsense mutations, out of frame deletions and insertions) either in germline (97%) or somatic (95%) events. The second includes VHL and WT1 and shows a difference between germline and somatic mutations, with most of germline mutations being missense (66% and 78%, respectively) while most somatic mutations are MLTPs (71% and 77%, respectively). Finally, the third group, p53, shows a majority of missense mutations either as germline (83%) or as somatic (80%) events. If we look at the level of missense mutations and study the distribution of transitions and transversions, we observe that for APC, WT1, and TP53, transitions are the most frequent germline and somatic mutations, while for VHL, transversions are the most frequent somatic events [Gallou et al., 1999]. It is clear that carcinogenic exposure is responsible for many tumors and specifically for some renal cell carcinoma (RCC). The study of a subgroup of patients (where the mutation is a transversion) for all these genes should be critical for the identification of toxic factors and the definition of professional risk.

GENETIC DISEASES DATABASES

Two LSDBs were created for genes specifically involved in genetic diseases. The development of the FBN1 LSDB required a modification in the structure of the UMD software to include clinical data [Collod et al., 1996]. This modification was subsequently added to most UMD softwares (Table 1). As for many genes the search of genotype/phenotype correlations is critical, we developed the "Clinical Data" table. To answer any situation, we designed a simple table including only two values: a so-called "symptom" value and a "severity" value. This table is linked to the "Mutation" table and for one mutation all clinical data can be included whatever their number and nature. In these conditions, this structure can be used for any gene/pathology without modification. The structure of the UMD databases is presented in Figure 1.

Furthermore, as the FBN1 gene encodes a protein highly repetitive containing four modules with homology to the human epidermal growth factor (EGF) precursor (EGF-like modules), 43 modules

TABLE 3. Distribution of Mutations of Germline and Somatic Mutations

	APC germline	APC somatic	MEN1 germline	p53 germline	p53 somatic	VHL germline	VHL somatic	WT1 germline	WT1 somatic
TOTAL	826	762	183	190	10,038	386	222	115	31
Frameshifts	572	455	88	11	782	83	136	10	15
Deletions	501	345	63	8	534	56	113	7	9
Insertions	71	110	25	3	248	27	23	3	6
Point mutations	252	306	94	171	8,765	302	85	105	16
Missense	23	38	54	158	8,035	253	64	90	7
Nonsense	229	268	40	13	730	49	21	15	9
G→A	10	12	14	10	1,045	16	7	12	2
G→A at CpG	0	3	5	50	1,380	39	3	16	0
C→T	52	61	12	11	858	32	6	22	3
C→T at CpG	106	100	18	43	1,170	40	5	44	9
A→T	14	13	2	6	255	3	3	0	0
A→G	3	4	1	12	824	14	3	2	1
A→C	1	0	0	1	127	11	3	0	0
T→G	4	5	3	2	247	12	3	2	0
T→C	1	5	7	6	317	52	11	5	0
T→A	7	7	3	3	272	11	6	1	0
C→A	21	20	5	5	273	9	10	4	1
C→G	22	14	6	5	309	28	8	3	0
G→T	10	60	9	11	1,265	24	9	1	0
G→C	1	2	9	6	423	11	8	3	0
Complex mutations	2	1	1	8	491	1	1	0	0
Transitions	68%	60%	61%	77%	64%	64%	41%	87%	94%
Transversions	32%	40%	39%	23%	36%	36%	59%	13%	6%

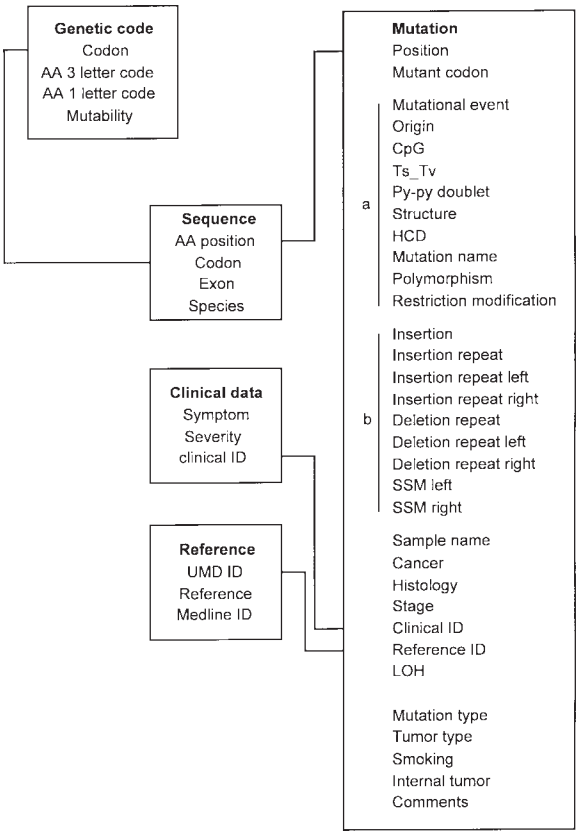

FIGURE 1. **Structure of the UMD software. Each box represents one of the tables and its name appears in bold. Note that some items from the "Mutation" table can be modified. Items directly calculated by the software are shown in a and b; furthermore, b contains items used for the analysis of repeated sequences involved in mutational events. AA = amino acid; Ts_Tv = transition or transversion; Py = pyrimidine; HCD = highly conserved domain; LOH = loss of heterozygosity; SSM = slipped strand mispairing.**

with homology to the EGF precursor and presenting a calcium-binding consensus sequence (cb EGF-like modules), eight modules homologous with transforming growth factor-β1 binding protein (eight cystein modules), we developed specific routines. The "Protein Menu" was created and it includes various routines to study the distribution of mutations in the different modules and aligns the amino acids of the consensus sequence for each module type (Table 2). The LDLR gene also contains repetitive domains and similar routines have been developed for this gene (Table 2). One of the goals of these LSDBs is to establish genotype/phenotype correlations and the collection of clinical information is critical. It would be worthless to collect as much information as possible if it is not correctly classified to give easy access to different levels of information. It is, thus, important that a community of experts in the field define a consensual submission form for clinical data to the LSDB curator, if possible before the LSDB is established.

If for each gene specific questions can be answered, a more general dimension is the understanding of molecular mechanisms involved in mutations. Some tools available in the UMD software give access to this level of analysis and we will use as an example the contribution of repeated sequences involved in deletions and insertions. As mentioned earlier, when the mutation is a deletion or an insertion, the UMD software searches automatically if flanking repeated sequences could be involved in the etiology of these mutations. Two routines, called "Deletion Analysis" and "Insertion Analysis" were used to produce the results shown in Table 4. A statistically significant difference is found between these two categories. For insertions, a repeated sequence is involved in 72.7% of cases vs. only 45.6% for deletions ($P < 0.001$). Another difference is observed in the length of the repeated sequence. The size of the repeated sequence is proportional to the size of the inserted sequence, while the size of the repeated sequence is independent of the deletion size (data not shown). This can be explained as follows: the repeated sequence is already present when a deletion process is involved, while it is the partial or total completion of an adjacent sequence that creates the repeated sequence in the insertion process.

EVOLUTION AND CORE DATABASES

We have come a long way since the creation of the first p53 database and software in 1994 [Cariello et al., 1994]. Many evolutions have taken place to produce the UMD software that is now available for Mac and PC to create LSDBs that can be used locally or accessible via the Internet. Since from the beginning we designed and made available many tools to analyze data, this software has now become a standard for locus database creation. To build this generic software choices had to be made, particularly in its structure. The UMD is today limited to the coding sequence and splice sites of a gene and cannot include large deletions or intronic mutations, which should be included for completeness as a flat file list. An evolution would be to include these mutations. However, this requires that the entire intronic sequence be known and the definition of a consensual wild-type intronic sequence. Today this is not possible for most genes. Another choice in the design of UMD was the development of mutation-oriented data-

TABLE 4. Results of the "Deletion Analysis" and "Insertion Analysis" Routines in 6 LSDBs

	APC	LDLR	MEN1	p53	VHL	WT1	Total
Deletions	856	71	63	990	190	16	2186
Del. with known sequence	646	71	63	567	189	16	1552
Del. with repeated sequence	288 (44.6%)	34 (47.9%)	37 (58.7%)	267 (47.1%)	76 (40.2%)	6 (37.5%)	708 (45.6%)
Insertions	181	24	25	268	51	10	559
Ins. with known sequence	139	23	22	99	40	10	333
Ins. with repeated sequence	119 (85.6%)	7 (30.4%)	20 (90.9%)	67 (67.7%)	22 (55%)	7 (70%)	242 (72.7%)

Deletion with known sequence = deletions for which the precise nucleotide position is known. Insertions with known sequence = insertions for which the precise nucleotide position of the insertion and the inserted sequence are known. Repeated sequence = two identical sequences found in the 5′ and the 3′ end of the deletion or the insertion.

bases which are useful for most molecular epidemiology studies. With the addition of the "Clinical Data" table came the question about unrestricted access to clinical data from specific patients. As countries have very different legislation, it is not easy to build a generic software including such data. In UMD, we chose the family level, meaning that we enter one mutational event for the family and associated clinical data can be stored as the number of affected patients/number of carriers that allows genotype/phenotype analysis. The hyperinsulinism community (HI) has asked for the creation of a link between the unique identifier in the SUR1-UMD database and a comprehensive and patient-specific clinical database. The access to the link will be restricted to people from the community. This approach can be a good evolution to store both levels of information.

The UMD software has been used to create different worldwide reference databases in the field (FBN1, VHL, and SUR1) and the question of restricted access has been frequently asked. It is not clear today for us if a worldwide reference database can have a restricted access, although this has some advantages.

In the near future, the entire sequence of the human genome will be available. The future development of new methods for the detection of point mutations such as the Chips technology will lead to a huge increase of new mutation detection. Although it is difficult to evaluate the number of mutations reported in the literature so far, it is also impossible to predict how many new mutations will be detected in the next 10 years. Nevertheless, several points are predictable. The rate of de novo mutation will never slow down or stop either for somatic or germline mutation. Furthermore, changes in our environment will lead to changes in the mutational events which modify our genome. Thus, the task of reporting and analyzing these mutations will be a major challenge in the future, especially if the presence or the identity of such mutations is linked to a therapeutic decision.

Finally, locus databases are essential associates of core databases such as HGMD [Cooper et al., 1998; Krawczak et al., 2000]. It is of great interest that data can be exchanged between locus databases and core databases because curators play a major role in controlling data from locus databases and can avoid the numerous errors found in publications. It will also be of great interest to facilitate the submission of mutations to locus databases using a common submission form [for review, see the HUGO Mutation Database Initiative (MDI); Cotton et al., 1998; http://ariel.ucs.unimelb.edu.au:80/~cotton/entry.htm].

REFERENCES

Antonarakis SE, the Nomenclature Working Group. 1998. Recommendations for a nomenclature system for human gene mutations. Hum Mutat 11:1–3. http://www.interscience.wiley.com/jpages/1059-7794/nomenclature.html.

Benson DA, Boguski MS, Lipman DJ, Ostell J. 1997. GenBank. Nucleic Acids Research 25:1–6.

Béroud C, Soussi T. 1996. APC gene: database of germline and somatic mutations in human tumors and cell lines. Nucleic Acids Res 24:121–124.

Béroud C, Soussi T. 1997. p53 and APC gene mutations: software and databases. Nucleic Acids Res 25:138.

Béroud C, Soussi T. 1998. p53 gene mutation: software and database. Nucleic Acids Res 26:200–204.

Béroud C, Verdier F, Soussi T. 1996. p53 gene mutation: software and database. Nucleic Acids Res 24:147–150.

Béroud C, Joly D, Gallou C, Staroz F, Orfanelli MT, Junien C. 1998. Software and database for the analysis of mutations in the VHL gene. Nucleic Acids Res 26:256–258.

Cariello NF, Cui L, Béroud C, Soussi T. 1994. Database and software for the analysis of mutations in the human p53 gene. Cancer Res 54:4454–4460.

Collod G, Béroud C, Soussi T, Junien C, Boileau C. 1996. Software and database for the analysis of mutations in the human FBN1 gene. Nucleic Acids Res 24:137–140.

Collod-Béroud G, Béroud C, Ades L, Black C, Boxer M, Brock DJ, Godfrey M, Hayward C, Karttunen L, Milewicz D, Peltonen L, Richards RI, Wang M, Junien C, Boileau C. 1997. Marfan database (second edition): software and database for the analysis of mutations in the human FBN1 gene. Nucleic Acids Res 25:147–150.

Collod-Béroud G, Béroud C, Ades L, Black C, Boxer M, Brock DJ, Holman KJ, de Paepe A, Francke U, Grau U, Hayward C, Klein HG, Liu W, Nuytinck L, Peltonen L, Alvarez Perez AB, Rantamaki T, Junien C, Boileau C. 1998. Marfan database, 3rd ed. New mutations and new routines for the software. Nucleic Acids Res 26:229–223.

Cooper DN, Ball EV, Krawczak M. 1998. The human gene mutation database. Nucleic Acids Res 26:285–287.

Cotton RG, McKusick V, Scriver CR. 1998. The HUGO mutation database initiative. Science 279:10–11.

den Dunnen JT, Antonarakis SE. 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat 15:7–12.

Edery P, Lyonnet S, Mulligan LM, Pelet A, Dow E, Abel L, Holder S, Nihoul-Fekete C, Ponder BAJ, Munnich A. 1994. Mutations of the RET proto-oncogene in Hirschsprung's disease. Nature 367:378–380.

Gallou C, Joly D, Méjean A, Starosz F, Martin N, Tarlet G, Orfanelli MT, Bouvier R, Droz D, Chrétien Y, Maréchal JM, Richard S, Junien C, Béroud C. 1999. Mutations of the VHL gene is sporadic renal cell carcinoma definition of a risk factor for VHL patients to develop an RCC. Hum Mutat 13:464–475.

Hofstra RMW, Landsvater RM, Ceccherini I, Stulp RP, Stelwagen T, Luo Y, Pasini B, Hoppener JWM, Ploos van Amstel HK, Romeo G, Lips CJM, Buys CHCM. 1994. A mutation in the RET proto-oncogene associated with multiple endocrine neoplasia type 2B and sporadic medullary thyroid carcinoma. Nature 367:375–376.

Jeanpierre C, Béroud C, Niaudet P, Junien C. 1998. Software and

database for the analysis of mutations in the human WT1 gene. Nucleic Acids Res 26:271–274.

Krawczak M, Ball EV, Fenton I, Stenson PD, Abeysinghe S, Thomas N, Cooper DN. 2000. Human Gene Mutation Database—a biomedical information and research resource. Hum Mutat 15:45–51.

Laurent-Puig P, Béroud C, Soussi T. 1998. APC gene: database of germline and somatic mutations in human tumors and cell lines. Nucleic Acids Res 26:269–270.

Mulligan LM, Kwok JBJ, Healey CS, Elsdon MJ, Eng C, Gardner E, Love DR, Mole SE, Moore JK, Papi L, Ponder MA, Telenius H, Tunnacliff A, Ponder BAJ. 1993. Germ-line mutations of the RET proto-oncogene in multiple endocrine neoplasia type 2A. Nature 363:458–460.

Olschwang S, Tiret A, Laurentpuig P, Muleris M, Parc R, Thomas G. 1993. Restriction of ocular fundus lesions to a specific subgroup of APC mutations in adenomatous polyposis coli patients. Cell 75:959–968.

Romeo G, Ronchetto P, Luo Y, Barone V, Seri M, Ceccherini I, Pasini B, Bocciardi R, Lerone M, Kaariainen H, Martucciello G. 1994. Point mutations affecting the tyrosine kinase domain of the RET proto-oncogene in Hirschsprung's disease. Nature 367:377–378.

Soussi T, Dehouche K, Béroud C. 2000. p53 Website and analysis of p53 gene mutations in human cancer: forging a link between epidemiology and carcinogenesis. Hum Mutat 15:105–113.

Spirio L, Olschwang S, Groden J, Robertson M, Samowitz W, Joslyn G, Gelbert L, Thliveris A, Carlson M, Otterud B, Lynch H, Watson P, Lynch P, Laurentpuig P, Burt R, Hughes J-P, Thomas G, Leppert M, White R. 1993. Alleles of the APC gene: an attenuated form of familial polyposis. Cell 75:951–957.

Varret M, Rabes JP, Collod-Béroud G, Junien C, Boileau C, Béroud C. 1997. Software and database for the analysis of mutations in the human LDL receptor gene. Nucleic Acids Res 25:172–180.

Varret M, Rabes JP, Thiart R, Kotze MJ, Baron H, Cenarro A, Descamps O, Ebhardt M, Hondelijn JC, Kostner GM, Miyake Y, Pocovi M, Schmidt H, Schuster H, Stuhrmann M, Yamamura T, Junien C, Béroud C, Boileau C. 1998. LDLR Database, 2nd ed. New additions to the database and the software, and results of the first molecular analysis. Nucleic Acids Res 26:248–252.

Xue F, Yu H, Maurer LH, Memoli VA, Nutile-McMenemey N, Schuster MK, Bowden DW, Mao J, Noll WW. 1994. Germline RET mutations in MEN 2A and FMTC and their detection by simple DNA diagnostic tests. Hum Mol Genet 3:635–638.