PAH Mutation Analysis Consortium Database: 1997. Prototype for relational locus-specific mutation databases

Piotr M. Nowacki, Susan Byck, Lynne Prevost and Charles R. Scriver*

The DeBelle Laboratory, McGill University-Montreal Children's Hospital Research Institute, 2300 Tupper Street, Montreal, Quebec H3H 1P3, Canada and Department of Human Genetics, McGill University, Montreal, Quebec, Canada

Received September 29, 1997; Accepted October 3, 1997

ABSTRACT

PAHdb **(http://www.mcgill.ca/pahdb) is a curated relational database (Fig. 1) of nucleotide variation in the human** ***PAH*** **cDNA (GenBank U49897). Among 328 different mutations by state (Fig. 2) the majority are rare mutations causing hyperphenylalaninemia (HPA) (OMIM 261600), the remainder are polymorphic variants without apparent effect on phenotype.** ***PAHdb*** **modules contain mutations, polymorphic haplotypes, genotype–phenotype correlations, expression analysis, sources of information and the reference sequence; the database also contains pages of clinical information and data on three ENU mouse orthologues of human HPA. Only six different mutations account for 60% of human HPA chromosomes worldwide, mutations stratify by population and geographic region, and the Oriental and Caucasian mutation sets are different (Fig. 3).** ***PAHdb*** **provides curated electronic publication and one third of its incoming reports are direct submissions. Each different mutation receives a systematic (nucleotide) name and a unique identifier (UID). Data are accessed both by a Newsletter and a search engine on the website; integrity of the database is ensured by keeping the curated template offline. There have been >6500 online interrogations of the website.**

Genetics is the study of heredity, biological variation and the mechanisms thereof. Genomics is the study of genomes and the genes they harbour. Medicine is the science (and art) of health maintenance and disease cure, alleviation and prevention. An abiding hope (and belief) is that the Human Genome Project will bring about a great instauration of 'genetic medicine', where genetic and medical thinking are fused. But in the abundance of information emerging in genetics and genomics and from the technologies they have spawned there arises a concern that the attendant knowledge could be obscured by the sheer mass of information. How then to record the essential information about variation in DNA, about mutations both disease-causing and neutral, and have it available to translate into genetic and medical knowledge? This is a goal of mutation databases to which the Human Genome Organization (HUGO) is committed (1).

THE PAH MUTATION DATABASE (http://www.mcgill.ca/pahdb)

This relatively large relational database (*PAHdb*) documents variation (mutations) in nucleotide sequence of the human cDNA for the phenylalanine hydroxylase gene [symbol *PAH*; GenBank U49897. Mutation numbering from the *PAHdb* annotated sequence begins with the A of ATG (codon 1) 473 nt into the cDNA sequence; nucleotide numbering in U49897 begins with the first 5′ base. NCBI has software to find the ATG codon and adjust numbering for nomenclature purposes.] and its enzyme (SWISS-PROT P00439); *PAHdb* records variation both disease-causing (notably those causing phenylketonuria, OMIM 261600) and neutral (polymorphic). *PAHdb* has been built as a prototype for other locus-specific mutation databases that will serve genetic medicine; it is already linked to centralized human genetic databases such as Human Genome Mutation Database (HGMD) (2) (http://www.cf.ac.uk/uwcm/mg/hgmd0.html); and Online Mendelian Inheritance in Man (OMIM) (http://www.ncbi.nlm.nih.gov/Omim). It could also be linked to interorganismal databases such as XREF (3). This, the third in a series of reports (4,5), describes continuing growth, use and relevance of *PAHdb*.

THE DATABASE: STRUCTURE, DEVELOPMENT AND MAINTENANCE

PAHdb has in excess of 4000 records in >60 data tables distributed in conceptual modules (Fig. 1). An overview of the design is the purpose of this section.

Documentation

A mutation database is a record of discovery and intellectual property and at the moment, the pace of mutation discovery far exceeds the capacity to publish the data by conventional means; an alternative is electronic publication, now in databases, eventually in journals. Peer review is always required, hence *PAHdb* is curated (edited) and under constant scrutiny by its

*To whom correspondence should be addressed. Tel: +1 514 934 4417; Fax: +1 514 934 4329; Email: mc77@musica.mcgill.ca

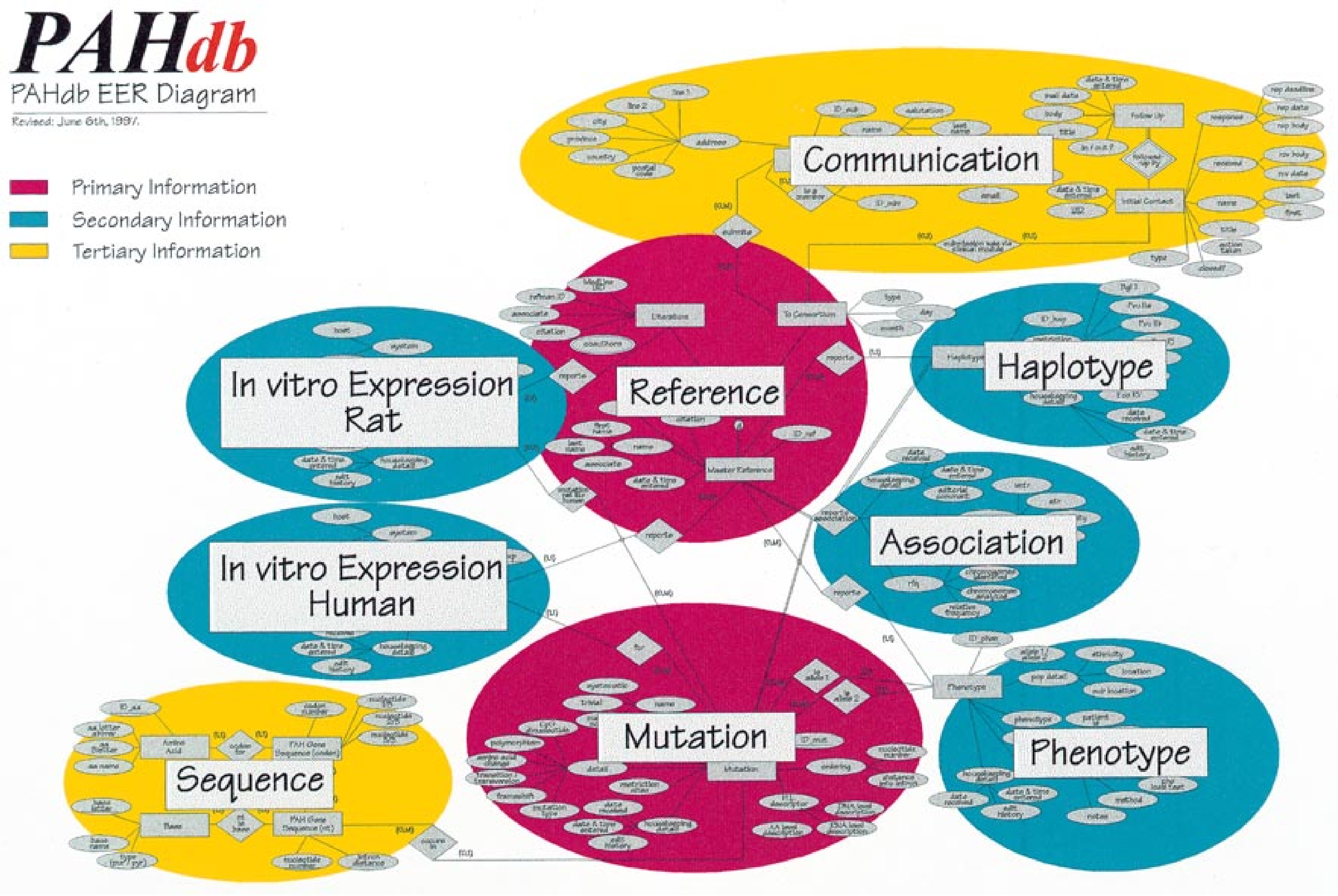


Figure 1. The entity relationship diagram for *PAHdb* showing modules containing attributes and descriptors of mutations.

Consortium members (and users). Curators will come and go; hence the database must be documented for continuity; meaning, its conceptual and physical design will always have a formal (evolving) record describing the flow of data, with instructions to users of its interfaces. The Document will be deposited as part of a thesis by one of us (PMN).

Design

Modules form the structure of *PAHdb* (Fig. 1). Each locus-specific mutation database will have a particular design and content but some degree of standardization is desirable if they are to be linked to centralized repositories and information is to be shared. Elsewhere we have proposed a way to achieve consensus through agreement on types of information (6). *Primary information* is standardized. It describes the mutation in biological terms using a sequence-based nomenclature. For example, the human M1V mutation in *PAH* (7) has three contexts: the human genome, a gene name and an accession number (GenBank) for the nucleotide sequence; thus: HSAP:*PAH*:U49897::M1V. Whereas the M1V mutation was initially reported by its trivial name it also has a systematic mutation name (c.1A→G, where 1 is the mutated nucleotide, A→G is the mutation, and c. refers to cDNA, the source of the annotated *PAH* sequence in the database); see S.E.Antonarakis at http://www.bio.net/hypermail/MUTATION/9705/0007.html for guidelines on mutation nomenclature. Accordingly, information on the M1V mutation is fully conveyed by: HSAP:*PAH*:U49897::c.1A→G; the trivial name (M1V) is a useful but not a necessary convention.

Secondary information covers attributes and descriptors of the mutation. For example, the association of the M1V mutation with its polymorphic haplotype background and the population and geographic region in which it is found; M1V was originally reported on *PAH* haplotype 2 and it occurs in Quebec (French Canadians) and France (French). The additional attributes and descriptors of a mutation are edited and listed in *PAHdb*. *Tertiary information* is that which is curatorial in nature, including the edit histories and comments; they are part of the record and include unique identifiers for the mutations. Under ideal conditions, patient (sample) identifiers would have been assigned to trace alleles (chromosomes).

Software

PAHdb currently uses Visual FoxPro 5.0 as its database management system (DBMS) and VFP, an object-oriented derivative of the original X-Base language, for programming and development.

Search engine

An offline copy of *PAHdb* is continuously edited and updated by the curators; edited copies are deposited at intervals on the server (http://www.mcgill.ca/pahdb) where it can be used. A search engine (West Wind Web Connection http://www.west-wind.com) gives Web users the opportunity to interrogate the database. Integrity is ensured because the original (edited) template is offline.

THE DATABASE: CONTENT AND UTILIZATION

Nomenclature and identity

Whereas the visual summary of *PAH* mutations (Fig. 2) uses their trivial names (8), the database and Newsletter provide the systematic (nucleotide) names, keyed to the 'Konecki' cDNA sequence (9). Unique identifiers (UIDs) have been given to each *PAH* mutation different by state.

Since *PAHdb* is a relational database it requires records to be linked on unique keys to keep tables properly related, and to avoid dangling tuples, as well as insertion and deletion anomalies. Use of the trivial mutation name alone is insufficient; for example, there could have been loss of data when the Y204C 'missense' mutation name was changed to the systematic name E6nt-96A→g recognizing that it is actually a splice mutation (10). Earlier versions of the DBMS (FoxPro 2.x) did not automatically enforce referential integrity but that difficulty has been overcome in Visual FoxPro and enhanced by the use of UIDs. On the other hand, the current system does not yet assign different UIDs to recurrent mutations (for example, R408W has one UID, number 121, but two origins in recurrent mutation) (11).

Alleles, chromosomes and UIDs

PAHdb records 12 279 'alleles'. To use the alternative term, 'chromosome', implies the source of the DNA sequence harbouring the mutation; a feature that should merit an identifier for each chromosome and the person of origin. *PAHdb* has not been able to take this step. Here, we use the term 'allele' mainly to implicate the possibility of different background nucleotide sequences in the *PAH* gene. We know that mutations identical by state can be different alleles; recurrent mutations are different alleles (e.g. R408W) (11). As knowledge emerges about the haplotypes harbouring mutations, and ultimately about the full genomic *PAH* sequences harbouring them, it should be possible to describe the relative degree of uniqueness of mutations.

There is a practical short-term problem: The record of 12 279 'alleles' is clearly not a record of that many independent chromosomes. The database uses reports, both published and submitted (see module). However a cumulative record for a given region or population does not obliterate the earlier (partial) reports. In the absence of assigning a UID for each chromosome (allele) at the initial report, *PAHdb* cannot document the actual number of independent alleles in the database. We had to deal with this deficiency in database design when calculating relative population frequencies of particular mutations (see below).

Mutations

On September 10, 1997 there were 328 different mutations (by state) in *PAHdb*; their distribution in the gene is shown in Figure 2. Since they were ascertained in probands with hyperphenylalaninemia (HPA), by definition those due to deficient phenylalanine hydroxylase function (12), the majority of these mutations are likely to be 'disease-causing', but not necessarily so. Although some are apparently neutral polymorphisms, notably the 'silent' mutations in codons and mutations deep in introns, it is too early to say categorically that all *PAH* polymorphisms are without any effect on *PAH* gene function.

PAHdb also classifies mutations by type (Fig. 2, inset). The majority (60%) are missense; to know whether they are disease-causing requires either expression analysis (see module in Fig. 1), other knowledge about the evolutionary importance of the affected residues or careful analysis of the genotype–phenotype relationship *in vivo*. The other *PAH* mutations behave largely as nulls; nonsense (6%) and splice (13%) mutations together are a minority in this gene; deletions (13%), with five exceptions, are all small (<20 bp in length) and most produce frameshifts; the few insertions (1%) are also small, with frameshifts.

Genotype–phenotype correlations

A meta-analysis (13) of 365 probands harbouring 73 different *PAH* mutations in 161 different genotypes assigned mutations to 'severe', 'intermediate' and 'mild' classes by their effect on plasma phenylalanine homeostasis. The *PAH* genotype does not sufficiently explain the mutant phenotype in all cases implying that other factors are involved in the emergent property of homeostasis (13,14). A copy of the raw data set is stored in a compressed file on the *PAHdb* web site.

Expression analysis and molecular modelling

A module (co-edited by P.J.Waters) dedicated to expression analysis records 58 expression studies on 36 different human *PAH* mutations in seven different host systems (55% are in COS) using 10 different vectors (15). Another module contains data on 22 different site-directed modifications of the rat gene (three corresponding to human alleles) and their effects on the rat enzyme (15), which up to now has been its best studied species.

In a recent communication to *PAHdb*, R.C.Stevens (Berkeley, CA) mentions success with human PAH enzyme crystallization leading to a description of its 3D structure; in which case there is the possibility of molecular modelling *in silico* in parallel with the corresponding *in vitro* expression analysis.

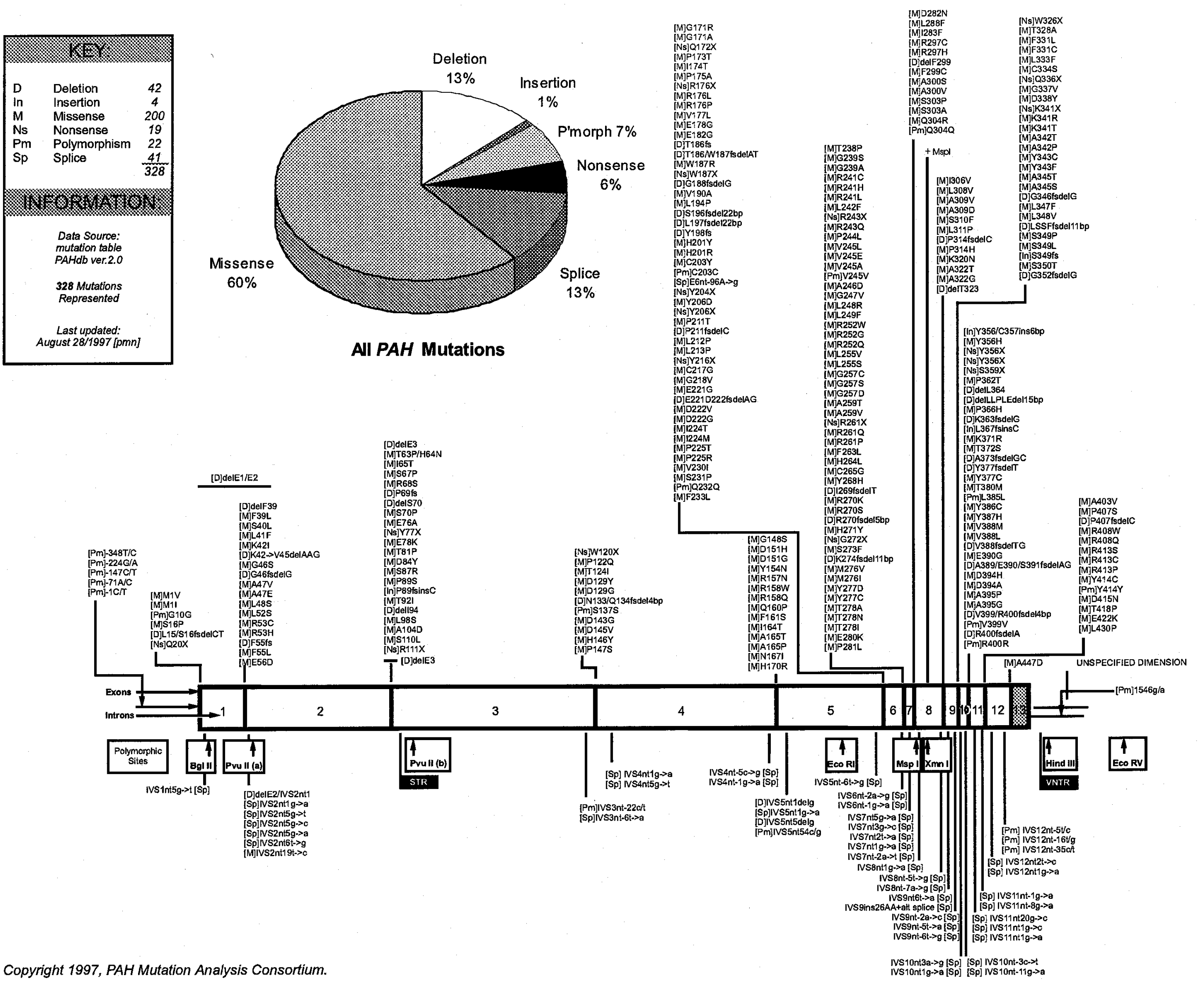
Relative frequencies of *PAH* mutations

From 340 articles and 180 unpublished reports currently in the relevant module, we extracted data on relative frequencies of *PAH* mutations. Mutations (stratified by geographic region and population) are not shared on European and Oriental chromosomes (Fig. 3), and only six different mutations account for 60% of the rare mutant haplotypes in the human population.

Polymorphic *PAH* haplotypes

The *PAH* locus is well endowed with known polymorphic markers (see Fig. 2); they include seven diallelic sites (RFLPs), two multiallelic markers (STR and VNTR), several silent SNPs in codons (e.g. Q232Q, V245V, L385L and Y414Y) and polymorphisms in the introns (e.g. IVS3nt-22c/t). From some of these markers, haplotypes are derived and associations with mutations analyzed.

Whereas the rare disease-producing *PAH* mutations are useful to study population histories (16), their presence is necessary for such studies and not all human populations harbour these mutations at useful frequencies. Instead, and under appropriate



Copyright 1997, PAH Mutation Analysis Consortium.

Figure 2. Structure of the *PAH* gene locus (~100 kb; chromosome 12q24.1) indicating relative positions of exons and size of introns. The complete genomic sequence has not yet been reported; *PAHdb* uses the cDNA sequence for systematic mutation names; mutation numbering begins with the A of ATG (codon 1, methionine). Mutations are shown by type (see key) and by trivial names. Polymorphic markers used for haplotypes are shown below the gene (open boxes, diallelic; shaded box, multiallelic); some known 'silent' polymorphisms (e.g. IVS3nt-22c/t and Q232Q) are indicated.

conditions, the polymorphic *PAH* markers and haplotypes can be used for this purpose (17).

Other features

The clinical page (http://www.mcgill.ca/pahdb/handout/handout.htm) has wide acceptance and use. Students consult it; patients use it to gain information and pose questions (which the curators answer) and physicians use it for teaching. *A mouse Pah gene page* describes the mutagenized (ENU) mouse counterpart and orthologous phenotype (co-editor J.D.McDonald Wichita State University); the ENU mouse strains are powerful allies to study the physiology of *PAH* mutation effects and a new treatment option (18). *The cDNA sequence* is annotated and all available information about intron sequences has been incorporated.

Visits and users

The Consortium Newsletter is distributed to 88 investigators in 28 countries. The counter on the website recorded 6431 visits over 429 days (up to September 11, 1997) from around the world (Fig. 4); 12% of visits originate from the .com domain; 10% from the .edu domain.

DISCUSSION

PAHdb is a relational database currently providing information on 328 different mutations at the *PAH* locus. A bias of ascertainment influences this record. First, the mutations were largely ascertained in probands with hyperphenylalaninemia; we do not know about the non-hyperphenylalaninemic population. Second, probands were largely ascertained in populations where newborn screening

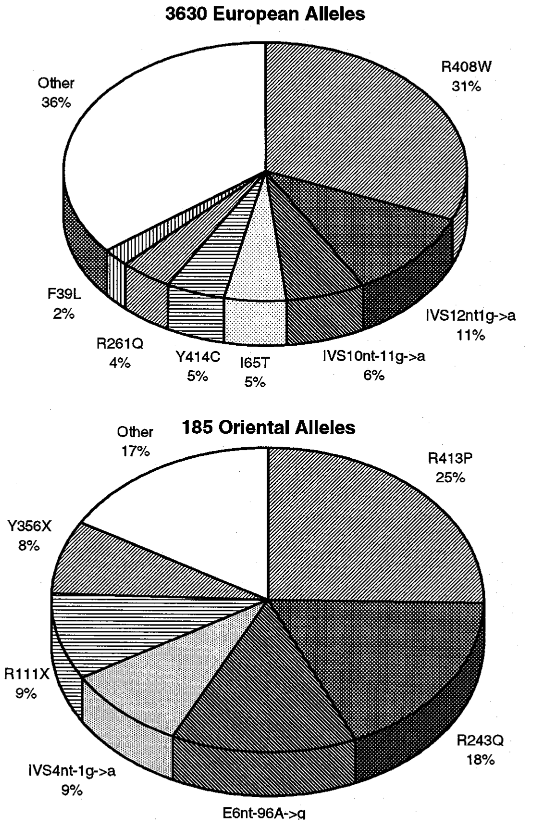


Figure 3. Europeans (Caucasians) and Orientals (from China, Korea and Japan) have different *PAH* mutation profiles.

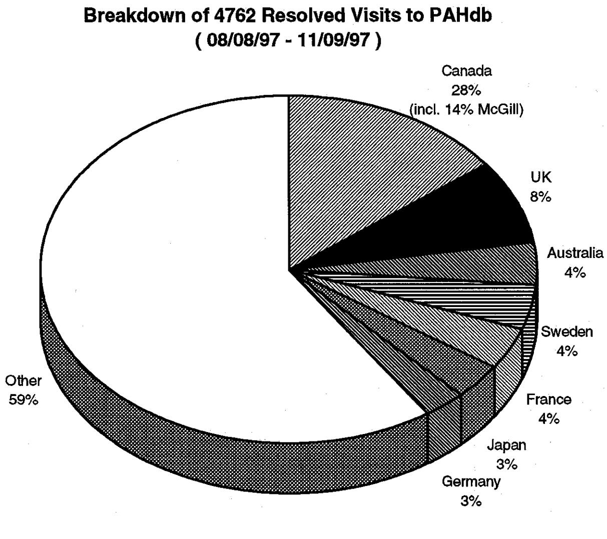


Figure 4. Visits (recorded) to the *PAHdb* website stratified by place of origin; 6431 visits during 429 days for which 4762 were resolved for the display here.

for hyperphenylalaninemia takes place as a public health practice; most Oriental and African populations, for example, are not screened. Accordingly the record of human *PAH* mutations is skewed.

PAHdb provides information that is both universal and relevant to the whole genome; and it is particular, being descriptive only of the *PAH* locus. The former is relevant to initiatives designed to survey the whole human genome. H.Lehvaslaiho (personal communication) lists eight fields that can be surveyed by SRS-based software in >20 different locus-specific mutation databases; they include species, genes, protein change, amino acid change, DNA change (both length, position and nucleotide) and mutation type. The locus-specific information is of interest to those attempting to relate a mutation to other features such as haplotype, phenotype, population structure and so on. The time has arrived to establish guidelines whereby genomic (central) databases and locus-specific databases can be linked to share information.

PAHdb has assigned unique identifiers (UIDs) to mutations. This practice has raised a problem yet to be resolved and requiring guidelines. Mutations on different chromosomes (alleles) may be identical-by-state but not necessarily identical by descent. For example, the R408W mutation (c.1222C→T) is identical-by-state on haplotypes 1.8 and 2.3 but not identical-by-descent because evidence indicates it is a recurrent mutation at a CpG site (11); the same is true for the E280K mutation (c.838G→A) on haplotypes 1.8 and 2.3 (19). Accordingly, each of the R408W and E280K mutations on their separate haplotypes could eventually receive a different UID. On the other hand the G218V mutation, which occurs on haplotypes 1 and 7 by virtue of a single intragenic recombination (20) and is therefore identical by descent, would have only a single UID.

PAHdb, which documents mutation types, shows they may differ among genes. For example, *PAH* has few protein-truncation mutations compared with *BRCA1* and few large deletions or insertions compared with *LDLR*. These differences reflect locus-specific features and will be informative about the fine structure of the human genome and the corresponding homologues in non-human species.

PAHdb documents relationships between genotypes and phenotypes. Accordingly, an interesting dimension of genetics, such as the function of a gene and parameters of related homeostasis, will emerge. This essential knowledge need not be lost in the mass of information emerging from the genome project and locus-specific mutation databases are one mechanism to retrieve and display it.

PAHdb contains intellectual property thus raising the issue of copyright. The data in our and other locus-specific mutation databases are freely shared with the broad community of scientists, educators, patients and others; the commercial sector is also a user of our database (Fig. 4). The *PAHdb* home page advises users that information in the database is copyrighted intellectual property; users should cite the source of their information in the appropriate manner.

ACKNOWLEDGEMENTS

We thank Max Cacchione for his assistance with annotation of the *PAH* nucleotide sequence, and Ken Hechtman for other *PAHdb* related tasks. *PAHdb* has been supported in part by the Medical Research Council (Canada), the Canadian Genetic Diseases Network (Network of Centers of Excellence), the Quebec Network of Applied Molecular Genetics of FRSQ, and The Interuniversity Institute for Population Research (IREP).

REFERENCES

1 Scriver,C.R., Cotton,R., Antonarakis,S. and McKusick,V. (1997) *Genome Digest*, **4**, 12–15.

2 Krawczak,M. and Cooper,D.N. (1995) *Nature*, **374**, 402

3 Bassett,D.E.J., Boguski,M.S., Spencer,F., Reeves,R., Su-hyon,K., Weaver,T. and Hieter,P. (1997) *Nature Genet.*, **15**, 339–344.

4 Hoang,L., Byck,S., Prevost,L., Scriver,C.R. and curators (1996) *Nucleic Acids Res.*, **24**, 127–131.

5 Nowacki,P., Byck,S., Prevost,L. and Scriver,C.R. (1997) *Nucleic Acids Res.*, **25**, 139–142.

6 Nowacki,P. and Scriver,C.R. (1997) http://ariel.ucs.unimelb.edu.au:80/~cotton/proposals.htm

7 John,S.W., Rozen,R., Laframboise,R., Laberge,C. and Scriver,C.R. (1989) *Am. J. Hum. Genet.*, **45**, 905–909.

8 Beutler,E., McKusick,V.A., Motolsky,A.G., Scriver,C.R. and Hutchinson,F. (1996) *Hum. Mut.*, **8**, 203–206.

9 Konecki,D.S., Wang,Y., Trefz,F.K., Lichter-Konecki,U. and Woo,S.L.C. (1992) *Biochemistry*, **31**, 8363–8368.

10 Ellingsen,S., Knappskog,P.M. and Eiken,H.G. (1997) *Hum. Mut.*, **9**, 88–90.

11 Byck,S., Morgan,K., Tyfield,L., Dworniczak,B. and Scriver,C.R. (1994) *Hum. Mol. Genet.*, **3**, 1675–1677.

12 Scriver,C.R., Kaufman,S., Eisensmith,E. and Woo,S.L.C. (1995) In Scriver,C.R., Beaudet,A.L., Sly,W.S. and Valle,D. (eds), *The Metabolic and Molecular Bases of Inherited Disease*. 7th Edition. McGraw Hill Book Co, New York. pp. 1015–1075.

13 Kayaalp,E., Treacy,E., Waters,P.J., Byck,S., Nowacki,P. and Scriver,C.R. (1998) *Am. J. Hum. Genet.*, in press.

14 Treacy,E.P., Delente,J.J., Elkas,G., Carter,K., Lambert,M., Waters,P. and Scriver,C.R. (1997) *Pediatric Res.*, in press.

15 Waters,J., Parniak,M.A., Nowacki,P. and Scriver,C.R. (1997) *Hum. Mut.*, **10**, in press (December issue).

16 Scriver,C.R., Byck,S., Prevost,L., Hoang,L. and PAH Mutation Analysis Consortium (1996) In *Variation in the Human Genome*. John Wiley and Sons, Chichester, UK. pp. 73–96.

17 Byck,S., Morgan,K., Blanc,L. and Scriver,C.R. (1996) *Am. J. Hum. Genet*., **59**, A33(142) (Abstract).

18 Sarkissian,C., Lee,K.C., Danagher,P., Leung,R., Fuller,M.A. and Scriver,C.R. (1996) *Am. J. Hum. Genet*., **59**, A207(1183) (Abstract).

19 Byck,S., Tyfield,L., Carter,K. and Scriver,C.R. (1997) *Hum. Mut.*, **9**, 316–321.

20 Carter,K.C., Byck,S., Waters,P.J., Richards,B., Nowacki,P.M., Laframboise,R., Lambert,M., Treacy,E. and Scriver,C.R. (1998) *Eur. J. Hum. Genet*., in press.