Editorial

Mutation databases on the web

Why do we need mutation databases on the web?

Over the last five years the dramatic growth in the number of mutations identified in single genes has rendered their publication in printed databases increasingly difficult (fig 1). The need for such databases is beyond dispute, since they provide an accurate and up to date resource for the large community of molecular geneticists, doctors, bioinformaticists, and others interested in human genetic variation.

Concurrent with this growth in the number of mutations, the development of electronic publishing, remotely accessible over the internet, has provided a medium in which the size of the data field, its complexity, and its physical location are irrelevant to its publication. Moreover, more recent developments in the electronic publication of data means that this information can be interrogated and updated in real time. This new medium has thus provided, in a timely manner, an ideal vehicle through which to publish interactive databases of mutations.

The historical development of the world wide web

Before 1990, access to the internet was mediated solely through text menus using the Gopher protocol or alternatively through the use of cryptic common lines. The subsequent development of HTML (hypertext mark up language) graphical user interfaces (GUIs) by CERN transformed the internet into an intuitive "point and click" resource using graphics and sound to connect physically remote sites through virtual links. Accessing the internet through a GUI interface, more commonly known as a world wide web browser (the client), provides a convenient method for the non-computer specialist to access and interrogate remote sites (servers).

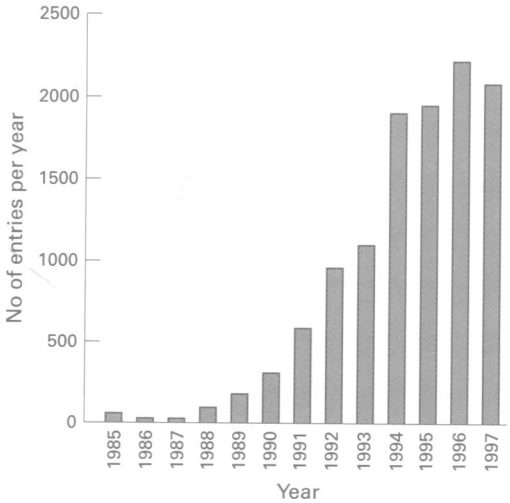

Figure 1 Number of mutants reported each year at the HGMD. The number for the whole of 1997 is extrapolated from half year results.

The technology underlying the web has been driven forward at breathtaking speed. Arguably the most important of the early advances was the development of interactivity between the client and server. This was originally mediated through server driven "CGI" scripts. These programs essentially worked by exploiting a characteristic of the UNIX operating system known as multitasking (a vital attribute for any server installed with the optimistic expectation of simultaneous access by multiple users). The CGI script would keep the browser waiting while performing a background task. Upon completion, the background task would then return a new page to the user.

The most recent development in internet technology is Java. Java is an object orientated programming language which is platform independent both at the source and binary level. It gives the same interactivity as CGI scripts, but is embedded in the web page. The Java programs execute on the client computer, and thus are both significantly faster to deliver/run and are also more flexible in their functionality than conventional CGI based scripts.

The perfect mutation database

Given unlimited time, finances, computer and network resources, the ideal mutation database would meet the requirement specification given in table 1. However, all software will fall far short of these naive objectives and will be bedevilled by compromises in order to bring a workable version of the product to shipment. In the rest of this article we shall attempt to review the current state of some of the remotely accessible WWW mutation databases and judge how they exploit the full range of technically possible features outlined above. Since it would be invidious to comment on or refer to our own mutation database site (HAMSTeRS), we leave it to readers to access it at http://europium.mrc.rpms.ac.uk and form their own opinions.

Analysis of the locus specific mutation databases

Approximately 40 databases are presently accessible over the world wide web. However, the quality and quantity of the data that they present is extraordinarily diverse. The URL (universal resource locators or WWW addresses) for these sites are given in table 2.

J Med Genet: first published as 10.1136/jmg.35.7.529 on 1 July 1998. Downloaded from http://jmg.bmj.com/ on October 5, 2022 by guest. Protected by copyright.

Table 1 Attributes of the ideal mutation database

Front end, of interest to the user: "Functional system requirements"	*Back end, of interest to the web master: "Non-functional system requirements"*
(1) Instant access	(9) Easy maintainability
(2) 100% accuracy of data	(10) 100% reliability (robustness)
(3) 100% timeliness (real time updating)	(11) 100% transparency to other databases
(4) Maximum user friendliness	
(5) Optimal data mining (eg Boolean search function)	
(6) Integration across the net to all other perfect databases	
(7) Perfect navigation within the database	
(8) A large quantity of reference information to put the mutations in context	

Most sites constitute a home page followed by a table of mutations and a limited number of references. The tables themselves contain data of varying quality ranging from as little as a simple download from OMIM to tables giving extensive quantities of biological information pertaining to all aspects of the respective disease state. Some sites have started to include internal links to anchors within the site and external links to pages outside the site. A broad quantity of background information on the significance of the mutations and the inclusion of submission/comment forms are also provided by the better sites. A very few of these sites also display a limited interactivity in addition to the functionalities described above. These search engines range in complexity from WAIS (wide area information searching) systems to full blown Boolean search algorithms.

An example of the best databases on the web can be found in the two Grap sites at Tromso and the NIH. The home page of the NIH site has a very user friendly clickable map linking the user to pages clearly explaining the region picked. Because these curators have created an imaginative facility, the user is excited to learn more about G protein coupled receptors. The only failing of this site, however, is that it contains a simple "excite" search of the database itself. Conversely the Tromso Grap site has a superb Boolean search engine and also includes the imaginative inclusion of "tiny grap", an algorithm that will return smaller (and thus more easily readable) tables from the database. Extensive external links between these sites would make this a single large but comprehensive site.

The NF1, Polycystic Kidney Disease, and BRCA1 databases were found to require a password in order to access the home page. Once penetrated, the visitor to the BRCA1 database encounters several pages of American style legal disclaimers. However, no one is refused access, and the real purpose of the protected entry may be to encourage scientists to deposit data ahead of publication in the refereed literature. The database is well presented with a search engine and links to other

Table 2 Universal Resource Locators (URL) of the web sites for global databases and individual disease states

Site	*URL*
EBI database	http://www2.ebi.ac.uk/mutations/integration
OMIM	http://www3.ncbi.nlm.nih.gov/omim
HGMD	http://www.cf.ac.uk/uwcm/mg/hgmd0.html
Mutation Database, Melbourne	http://ariel.ucs.unimelb.edu.au:80/~cotton/mut_database.htm
Cystic Fibrosis	http://www.genet.sickkids.on.ca/cftr/
Haemophilia A	http://europium.mrc.rpms.ac.uk
Phenylketonuria	http://blizzard.cc.mcgill.ca/pahdb/
Von Willebrand disease	http://mmg2.im.med.umich.edu/vWF/
HPRT1	http://sunsite.unc.edu/dnam/des_hprt.html
Cholinesterases	http://ensam.inra.fr/cholinesterase
Polycystic kidney disease	http://medoc.gdb.org/pkd/
Globins	http://globin.cse.psu.edu
Mitochondrial	http://www.gen.emory.edu/mitomap.html
Ataxia telangiectasia	http://www.med.jhu.edu/ataxia/mutate.htm
Skin Disease Mutation Database	http://www.bme.unc.edu/sdmdSearch.html
p53	http://www.mayo.edu/research/papers/P53%20Mutations/
Adenomatous polyposis coli	http://www.mayo.edu/research/papers/P53%20Mutations/
Androgen receptor	http://www.mcgill.ca/androgendb/
G protein coupled receptors, Tromso	http://www-grap.fagmed.uit.no/GRAP/homepage.html
G protein coupled receptors, NIH	http://mgddk1.niddk.nih.gov:8000/MutationAnalysis.html
G protein coupled receptors, Heidelberg	http://swift.embl-heidelberg.de/7tm/mutants/mutants.html
Acid alpha-glucosidase	http://www.eur.nl/FGG/CH1/glucosidase.html
L1 cell adhesion molecule	http://dnalab-www.uia.ac.be/dnalab/l1.html
Duchenne/Becker muscular dystrophy	http://ruly70.MedFac.LeidenUniv.nl/~duchenne/
Collagen	http://www.le.ac.uk/depts/ge/collagen/collagen.html
OTC	http://www.peds.umn.edu/otc/
X linked agammaglobulinaemia	http://www.helsinki.fi/science/signal/btkbase.html
Fanconi anaemia mutations	http://www.rockefeller.edu/fanconi/mutate
Mutation spectra database for bacterial and mammalian genes	http://info.med.yale.edu/mutbase/
Mucopolysaccharidosis disorders	http://www.peds.umn.edu/gene/
Emery-Dreifuss muscular dystrophy	http://www.path.cam.ac.uk/emd/
PAX6	http://www.hgu.mrc.ac.uk/Softdata/PAX6/
Ataxia telangiectasia	http://www.vmmc.org/vmrc/atm.htm
Tuberous sclerosis mutation database	http://www.cf.ac.uk/uwcm/mg/tsc_db/
Retinal degeneration, slow database	http://mol.ophth.uiowa.edu/MOL_WWW/RDStab.html
NF1 consortium	http://www.clam.com/nf/nf1gene/
Rhodopsin	http://mol.ophth.uiowa.edu/MOL_WWW/Rhotab.html
Hexosaminidase A	http://www.debelle.mcgill.ca/hexa/
MPS mutation	http://www.peds.umn.edu/centers/gene/

J Med Genet: first published as 10.1136/jmg.35.7.529 on 1 July 1998. Downloaded from http://jmg.bmj.com/ on October 5, 2022 by guest. Protected by copyright.

sites. Limited information on the structure and function of the normal gene product necessarily limits the usefulness of this site at present.

Personal thoughts on the locus specific databases

Most of the databases lack an imaginative exploitation of the dynamic interactivity of the web that is the hallmark of electronic publication and are thus little more than literal conversions of the printed databases. This is particularly damning when it is considered that CGI functionalities predated the establishment of most of these databases. However, there are several very well constructed sites and the standard seems to be improving.

It should be considered a minimum that each database should have a Boolean search algorithm. Ideally this interactivity should be "fuzzy", that is, it should accept and account for mistakes. Java applets also remain largely unused. Such applets could be coded to return the bioinformatics of the database in real time. In fact, the Java language contains numerous graphics library functions tailored to this very job. Multimedia applications, which could be used to broadcast complex biological information in an easily digestible format, are also unused by any site.

The last two years have also seen the development of VRML (virtual reality modelling language) enabled browsers that allow the user to view and manipulate 3D scenes in real time. This application is begging for use in the visualisation of protein structures within mutation sites to illustrate the effects of substitutions at the protein level. Finally, one of the advantages of the web is that it is remotely accessible by any user, including MDs, students, and sufferers of the disease state. We would suggest that each site should certainly address these communities by having a full explanation of the background of each disease state.

Core mutation databases

The proposition of a GDB for each chromosome, or a protein data bank for each protein fold, illustrates the requirement for a comprehensive and up to date core database consolidating the individual locus specific databases. Furthermore, of the 7006 missense/nonsense mutants identified to date, 85% occur in genes with fewer than 25 mutants per gene (fig 2) and new genes with mutations causing human disease are being published every week. Since it would be impractical to publish and access up to 50 000 individual web sites, the central database should act as a reservoir for all these rarer mutants.

These needs have been wholly or partially identified by four competing groups, that is, the mutation databases at Melbourne, the European Bioinformatics Institute Mutation Database, OMIM (Online Mendelian Inheritance in Man), and Cardiff's HGMD (Human Gene Mutation Database). The "non-functional" system requirements for a central database will be subtly different from those of the locus specific databases in order to account for the

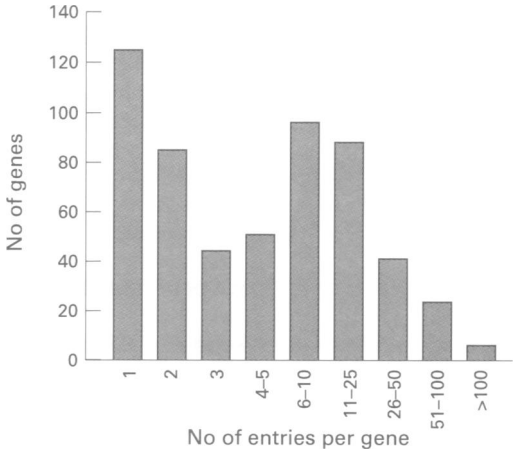

Figure 2 Frequency distribution of missense and nonsense mutations per gene (data from HGMD).

breadth of information contained therein. The main problem of knowledge management behind these non-functional system requirements is how to integrate transparently the differing fields of information reported by the many different researchers so that it is accessible to the whole community interested in variation in the genetic code. Less complex functional system requirements can be identified as: (1) how to provide simple access to such a varied database; (2) how to cross reference individual mutations in comparisons between evolutionary related proteins; (3) what the quality and quantity of background information for each gene will be; (4) how the database should link to the outside world; (5) how the bioinformatics of the database should be extracted and presented. (With such a large quantity of information available, mathematical modelling of parameters within the database should further elucidate the molecular mechanisms underlying DNA mutation and repair.)

The Mutation Database at Melbourne, though it has so far published two papers and met three times in exotic locations around the world, has not yet come up with a single line of code so it remains a purely hypothetical database. However, this situation may change after the fourth meeting of the Mutation Database Committee of HUGO in Baltimore on 27-28 October 1997. See website for details.

The EBI database at Hinkston Hall in Cambridge represents a currently active central database. The strategy behind this site is to integrate all the locus specific databases into a single coherent resource. The mutation data exported from locus specific sites may be searched by Boolean logic and this functionality is a significant advance on the simple tables presented by the majority of the locus specific sites. However, since the data originate from different sources, it contains different parameters for each gene and consequently the search fields also differ for each gene. Though explanations for each field can be accessed through a hyperlink, the resultant confusion generated by this system requirement obscures the synergistic effects that a core database should exploit. The site also lacks a bioinformatical analysis of the spectrum of single genes

J Med Genet: first published as 10.1136/jmg.35.7.529 on 1 July 1998. Downloaded from http://jmg.bmj.com/ on October 5, 2022 by guest. Protected by copyright.

or the human genome. The background for each gene is also limited, though this failing is abrogated by extensive connections to the locus specific databases. This site, surprisingly, has no capacity or stated ambition to accept novel mutants and relies solely on submissions from the locus specific databases and OMIM. On the plus side, this site has ambitious links to several dozen related resources, though as yet only a few of these appear to work. However, we believe that the computational strategy that the curators have adopted has established a solid foundation to overcome these minor problems.

An approach to the central database problem has been arrived at organically at the HGMD. Since 1988, Cooper and Krawczack have been compiling a comprehensive listing of all published mutations in order to elucidate the molecular mechanisms underlying mutations. The collection of data is achieved by computer and manual searches of the available publications. This database represents an extraordinarily comprehensive list of all reported mutants. However, it is limited to one entry and reference per mutation, thus excluding information about frequency, ethnogeographical origin, and phenotype-genotype variation. The site uses a WAIS "fuzzy" key word search which returns HTML links to each gene table. These pages contain data for missense/nonsense, splicing, and regulatory mutations and small insertions/small deletions and in most cases the cDNA sequence of the gene in question. These tables are themselves hyperlinked to OMIM, the GDB, and Medline. The site also gives regular updates of the bioinformatics of the database and extensive links outside the site. However, no Boolean search functions and no cross links between genes or from genes to structural information are as yet provided.

OMIM, the online version of McKusick's book, is the classic resource for all human inherited disease, containing frequently updated phenotype and genotype information. Fuzzy search terms and external links enable the user rapidly to collate available information and access primary publications painlessly. Mutations are simply listed in sequence by year of publication, though the list of mutations for each gene is by no means complete. McKusick is also collaborating with the projected Mutation Database, Melbourne.

Core databases: competition or cooperation?

OMIM remains the great core resource for human genetics and the Melbourne database is still a set of aspirations in the minds of its promoters. More interestingly, there is at present a competitive battle for users being fought between HGMD and EBI. Some two years after their establishment, HGMD appears to have stolen a march on the EBI since its list of mutations and working links is far stronger than that of EBI. On the other hand, EBI is far in front in terms of search functions. Whether Dawkinsian selection will drive one or more of these memes to extinction or they will continue to coexist in the ever evolving complexity of internet ecology, time alone will tell. Thankfully for the user, these differences are at best an inconvenience, since by using hyperlinked bookmarks seemless jumps can be made between the two sites.

Future development of mutation databases

What is the future of locus specific databases in an age where the core databases are growing rapidly in comprehensiveness and competence. In the first analysis it looks as though locus specific sites could become anachronisms or mere way stations to feed mutants to the core database. However, to consign the expertise behind these sites to such a menial function would be foolhardy. One answer to this dilemma is to integrate the expertise of the locus specific site curators into the core databases. Since these sites are by definition remotely accessible then they may also be remotely updated. By the judicious use of file permission flags it would be possible to assign sections of the core database to the locus specific site curators. This strategy would also have the additional benefits of removing a substantial workload from the core data base curators and engender a more democratic atmosphere among the community it serves.

It is clear that the number of individual mutations reported per year has reached a plateau (fig 1) which may correspond to the maximum fundable activity in mutation detection across the scientific community. It remains to be seen whether this trend will resolve as a decline in the number of mutants reported or whether new cheaper mutation detection technology will lead to a further upswing. Several groups are now involved in the analysis of whole national populations of disease states which, when reported, will cause large one off surges of new data. Even at present rates of accretion, the sheer mass of accumulated mutational information requires a major strategic rethink of the whole database project. Initially, submission methods must be automated and coded to include error checking mechanisms for each mutant/gene. Additionally the emphasis of the databases should evolve from lists of and searches for single mutants to bioinformatic analysis incorporating the structural databases and information from model organisms and the genome projects.

Biological significance of mutation data

Having gathered information on tens of thousands of mutations in hundreds of different genes what use can it be put to? At the level of a single gene, for example haemoglobin, the historical accumulation of mutation data has contributed enormously to the detailed mechanistic understanding of haemoglobin's function as an oxygen carrier. At the level of phenotype variation, it remains puzzling why some sickle disease patients are severely affected with progressive multiorgan damage and early death while others are suffering only mildly from

J Med Genet: first published as 10.1136/jmg.35.7.529 on 1 July 1998. Downloaded from http://jmg.bmj.com/ on October 5, 2022 by guest. Protected by copyright.

occasional crises. This opens up new questions about the influence of other genes on the phenotype, for example, those influencing haemoglobin F levels.

For other loci, such as BRCA1, even the mechanism connecting mutants to a cancer prone phenotype is currently obscure. Where a group of genes encodes similar proteins, for example the vitamin K dependent coagulation enzymes, inferences may be drawn from the mutational effects in one gene to the effect of mutation at a homologous residue in another member of the gene family.

By comparing mutational spectra across the whole range of human disease, inferences can be made about mechanisms operating at mutation hotspots.

It remains to be seen whether computer based analysis of the data in the global databases will yield new insights not already apparent from intensive manual scrutiny of the same data.[1]

Conclusion

These comments and proposals will already be out of date by the time this editorial is published, so fire up your web browser, enter net world, and watch it happening.

ADAM I WACEY
Thrombosis Research Institute, Emmanuel Kay Building, Manresa Road, London SW3 6LR, UK

EDWARD G D TUDDENHAM
MRC Clinical Sciences Centre, Haemostasis Research Group, Hammersmith Hospital, London W12 0NN, UK

1 Cooper DN, Krawczak M. *Human gene mutation*. Oxford: Bios Scientific Publishers, 1993.

J Med Genet: first published as 10.1136/jmg.35.7.529 on 1 July 1998. Downloaded from http://jmg.bmj.com/ on October 5, 2022 by guest. Protected by copyright.