

Biology's name game

The confused nomenclature of genetics is blighting the field — some genes have multiple names whereas unrelated genes often share a common moniker. Helen Pearson examines attempts to bring order to the chaos.

Fruitflies have *armadillo*, mice have *β-catenin*. Both of these genes produce proteins that influence early embryonic development, and their DNA sequences are so similar that *Drosophila* and mouse geneticists agreed several years ago that they are basically one and the same. Yet the two names persist.

Fly geneticists, with their penchant for extravagant gene names, regard *armadillo* as a perfect label for a gene that, when defective, gives *Drosophila* embryos an armour-plated appearance. Mouse geneticists, meanwhile, see no reason to rename a gene that clearly belongs to the *catenin* family.

Genes with multiple aliases seem to be the rule, rather than the exception, whereas genes that have no functional relationship with each other can often bear the same names. As biologists strive to make sense of the growing wealth of genomic information, this messy nomenclature is becoming a bugbear. “You’re

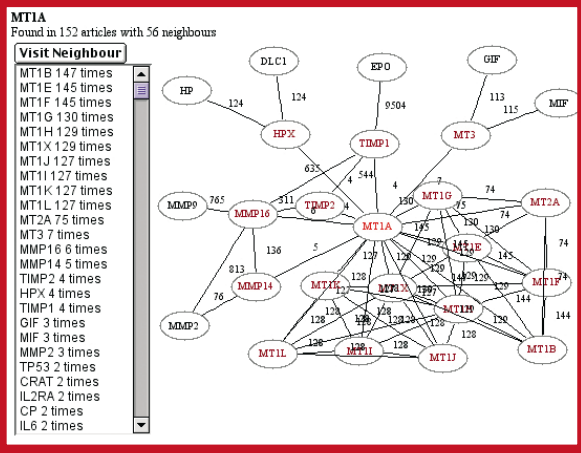
finding the closest relative of a gene is in a different organism,” says Judith Blake, who works on the Mouse Genome Database at the Jackson Laboratory in Bar Harbor, Maine. Unfortunately, gene names are proving a hindrance, rather than a help, in making these connections.

Everyone wants to be able to associate related genes with one another in the databases. But how should this be done? Attempts to impose standard names across the board are meeting stiff resistance, and approaches that would give genes unique ID numbers seem unlikely to take off unless journals enforce the system. But a coalition of leading geneticists may have the answer. The Gene Ontology (GO) Consortium is sidestepping the naming issue by developing ‘controlled vocabularies’. These will allow software to scan the genomic databases and link related genes to one another using terms that consistently describe their functions, regardless of what the genes are called.

Identity crisis

The extent of the nomenclature problem is illustrated by the work of a team led by Eivind Hovig at the Norwegian Radium Hospital in Oslo. Hovig and his colleagues are testing software designed to search for biological associations between genes based on their co-occurrence in the abstracts of published papers¹. By 1 May, their ongoing automated scan had examined more than

Genetic disorder: Eivind Hovig's studies have illustrated the naming confusion, including the multiple genes known as *MT1* (below).



Biologists would rather share their toothbrush than share a gene name. **Michael Ashburner**

10 million records in the Medline literature database, identifying 22,008 distinct human genes. Of these, 10,352 had more than one name. One gene, officially designated as *SELL*, or *selectin L*, which controls cell adhesion during immune responses, had 15 aliases.

Even more confusing, 4,257 abbreviated names were used to refer to more than one gene. Top of the list was *MT1*, used to describe at least 11 members of a cluster of genes encoding small proteins that bind to metal ions.

But at least those genes are structurally related — which is more than can be said for the five unconnected genes that have at some point been referred to as *PAP*, involved in processes ranging from cell differentiation to inflammation of the pancreas.

Between organisms, the problem multiplies. Start reading any genetics review paper and you are likely to be juggling the names preferred for genes in different species within a few paragraphs. For example, the yeast homologue of the human gene *PMS1*, which codes for a DNA repair protein, is called *PMS2*; whereas yeast *PMS1* corresponds to human *PMS2*.

In an effort to address the problem, researchers working on the human and mouse genome projects have nomenclature



Going strong: Midori Harris coordinates the GO project from the European Bioinformatics Institute.

committees that are collaborating with one another and encouraging their respective communities to adhere to agreed names. But as researchers from a range of backgrounds jump on the genomics bandwagon, consensus is proving more difficult to obtain. "It's become a much bigger enterprise," says Sue Povey of University College London, who heads the Human Genome Organisation's Gene Nomenclature Committee.

Getting a consensus on gene names for the full range of important laboratory organisms could well be impossible. Used to working in relative isolation, researchers studying different species have grown attached to their respective gene-naming traditions. *Drosophila* geneticists, for instance, delight in using colourful names — such as *hedgehog*, which produces a signalling protein involved in a range of developmental processes, and *lost in space*, which guides the growth of neurons — and they have no intention of letting other geneticists spoil their fun. Two international nomenclature workshops, held in 1997 (ref. 2) and 1999 (ref. 3), concluded that, apart from between mammals, attempting to standardize gene names across species was pointless.

"Biologists would rather share their toothbrush than share a gene name," says Michael Ashburner, joint head of the European Bioinformatics Institute (EBI) at Hinxton near Cambridge, and one of GO's founders. "Gene nomenclature is beyond redemption."

There is also a realization that genes may have several functions as they are expressed at different times during development or are transcribed in different ways — so names based on known functions may turn out to

be misleading. "It'll be many years before we can agree on a set of names," says Mark Boguski, senior vice-president of research and development at Rosetta Inpharmatics, a company in Kirkland, Washington, that specializes in exploiting genomic information.

Genetics by numbers

One possible solution is to identify individual genes by unique ID numbers, and rely on database curators to provide links between related genes. The Mouse Genome Database already does this, assigning each gene an ID and listing all sequences deposited in the GenBank database that are related to it. Every entry in GenBank also has an accession number. But although journals such as *Nature*, *Nature Genetics* and *Science* demand that their authors list GenBank accession numbers in papers that describe a gene for the first time, they seem unlikely to move rapidly towards a system of enforcing the use of gene IDs or GenBank numbers at each mention of a gene. And without such a system, gene numbers seem unlikely to solve the nomenclature problem.

This is where GO comes in. Rather than trying to impose a standardized system for gene names or numbers, GO's members are developing agreed vocabularies to describe molecular functions, biological processes and cellular components^{4,5}. Using these terms, it becomes possible to link related genes irrespective of their muddled nomenclature.

For example, subunits of the transcription factor TFIIA, a protein that regulates gene expression, are encoded by genes with different names in *Drosophila* and yeast, but which share the same GO terms. Describing their molecular function as 'general RNA polymerase II transcription factor' means that the fruitfly genes *TFIIA-L*, *TFIIA-S* and *TFIIA-S-2* can be readily associated with their yeast functional equivalents, *TOA1* and *TOA2*.

GO was started by the curators of some of the main model organism genome databases — the *Drosophila* community's FlyBase, the *Saccharomyces* Genome Database, which is

used by yeast geneticists, and the Mouse Genome Database. It has since been joined by the curators of the database for the plant *Arabidopsis*, and WormBase, which serves researchers working on the nematode *Caenorhabditis elegans*.

Curators of the databases assign GO terms to individual genes and their products, and the information is fed back to a central GO database, maintained on servers at Stanford University in California. Ultimately, the consortium aims to produce a fully searchable database of terms that explain the function of genes in all organisms.

Deciding on appropriate terms has involved much discussion. For example, the polysaccharide chitin is a component of cell walls in yeast, but in *Drosophila* it is found in the insects' external cuticle. As a result, the GO term 'chitin metabolism', used to refer to genes involved in chitin production, now has two daughter terms, 'cuticle chitin metabolism' and 'cell wall chitin metabolism', to satisfy the requirements of fruitfly and yeast geneticists. "The vocabularies are dynamic — a work in progress," says Midori Harris, who in April was appointed as GO's first full-time coordinator, based at the EBI.

GO is also sufficiently flexible to accommodate synonyms where researchers commonly use two terms to refer to the same process — for example cell division and cytokinesis. Instances where the same name is used to refer to different processes by different research communities are harder to disentangle, says Harris. To a yeast geneticist, 'mating' means the fusion of two cells; whereas researchers working on mice would use it in the more conventional sense. GO's solution has been to adopt the term 'mating (*sensu Saccharomyces*)' — "doing it in the same way as budding yeast", explains Harris.

The GO system is rapidly gaining popularity, and GO terms featured in recent papers describing the *Drosophila*⁶ and human⁷ genomes, and a comprehensive library of mouse genes⁸.

The project has just received a major boost in the form of a three-year, \$5-million grant from the US National Human Genome Research Institute in Bethesda, Maryland, and has some enthusiastic converts. "We will stop purchasing products that don't use GO," says Ken Fasman, Boston-based global head of R&D Informatics with the drugs company AstraZeneca.

GO, it seems, is all systems go. ■

Helen Pearson works in *Nature's* science writing team.

- Jenssen, T.-K., Lægreid, A., Komorowski, J. & Hovig, E. *Nature Genet.* **28**, 21–28 (2001).
- Blake, J. A. *et al.* *Genomics* **45**, 464–468 (1997).
- White, J. A. *et al.* *Genomics* **62**, 320–323 (1999).
- Ashburner, M. *et al.* *Nature Genet.* **25**, 25–29 (2000).
- Ashburner, M. *et al.* *Genome Res.* (in the press).
- Adams, M. D. *et al.* *Science* **287**, 2185–2195 (2000).
- Venter, J. C. *et al.* *Science* **291**, 1304–1351 (2001).
- Kawai, J. *et al.* *Nature* **409**, 685–690 (2001).

♦ <http://www.geneontology.org>

Attempting to standardize gene names across species is pointless.