Getting more from your sequences using the web

http://www.genetics.nature.com/gazing/

The Internet and, in particular, the web browser have freed us from the burdens of installing, maintaining and upgrading special software and databases for research. We provide direct links to web-based tools and a brief description of their use in a web-article (http://www.genetics.nature.com/gazing/). These tools allow researchers to retrieve sequence information from databases, search for homologues to a sequence, explore protein family relationships and predict structure from sequence.

Retrieving information is no longer limited to querying isolated databases because hypertext links interconnect databases, allowing rapid navigation between many sources. A sequence name can therefore lead to successive retrievals of annotations, publications, sequence relatives, conserved motifs, structures and genetic data.

The rapid growth of the general sequence databanks means that similarity searches are often more effective when limited to subsets. Indeed, sharpening the focus of a search both reduces the size of the output and lowers the background of chance hits. In some cases, organism-specific databases can be searched but the location of these resources is often unknown to potential users or is hard to find. We have collected direct links to organism-specific search engines to address this problem. We also make available a service for comparing a sequence of interest to the user's own database.

Whereas sequence databanks are increasing in size and redundancy, the number of protein families has been leveling off. This increases the value of family-specific databases, both for searching sensitivity and for predicting structure and function. Protein family features in a sequence of interest can be efficiently identified by searching against any of several family-specific databases. Conversely, regions of sequence similarity characteristic of a family can be used to detect more distant homologues in the sequence databanks. These searches can be launched successively by filling out a single form.

To get the most from an alignment, informative displays are essential. To illustrate the degree of conservation in an alignment, there are web tools that display aligned residues with colours or boxes or as stacks of residue letters. Other tools convert alignments to evolutionary trees, which are valuable for discerning subfamily relationships. A variety of structural features can be predicted by analysing single sequences: compositionally biased segments, coiled-coil regions, internal repeats, transmembrane-spanning segments and secondary structural elements. For proteins that have structures available, the web makes accessible fold classifications and direct visual comparisons between related structures.

Databases are constantly being updated, so how can you keep up? Register your sequence with an alerting service, and it will inform you by e-mail of relevant discoveries such as its mapped position, the discovery of a new homologue or that a genome centre has scooped you by sequencing a random clone. □

Elizabeth A. Greene & Steven Henikoff
The Fred Hutchinson Cancer Research Center, Seattle, USA. e-mail: eagreene@fhcrc.org

Networking nomenclature

http://genetics.nature.com/nomen/nomen_article.html

For many scientists, the naming of the gene upon its discovery is as important as the naming of a new baby—perhaps more so. When parents name a child, they take into account many factors, including family ancestry, the appearance and perceived (or hoped for) character of the child, and their own preferences. In contrast, there are few parents who worry about whether the name will make their child uniquely identifiable in school registers, and later in telephone directories and other name databases—otherwise, we'd have fewer 'John Smiths' and more names like 'Zowie Bowie' and 'Heavenly Haraani Tiger Lily'. When naming a gene, similar factors may come into play, but the importance of a unique name should not be underestimated. Recognizing the difficulties in arriving at an appropriate gene name, we have provided an article (http://genetics.nature.com/nomen/nomen_article.html) which gives an overview of the need for a nomenclature system, with live links to various nomenclature websites, including those for different species.

The problems now faced by nomenclature committees are of huge proportions. Roughly 10% of the human genes have been cloned and named, with the remainder predicted to be with us by 2005. Over the next 5–10 years, we are likely to have identified (and will thus require names for) all the genes of *Caenorhabditis elegans*, *Arabadopsis thaliana*, *Mus musculus*, *Drosophila melanogaster*, *Danio rerio* (zebrafish), rat, chicken, agriculturally important plants such as rice (and maybe wheat and barley) and probably the rat, sheep, goat, cow and horse.

In the best of all possible worlds, (i) orthologous genes should be named similarly across species (assuming that orthologues can be identified unequivocally), (ii) homologues should be named somewhat similarly, and (iii) the gene product (mRNA, cDNA and protein) should have the same name as the gene. Although there are a number of possible ways to group and name genes, many are context-dependent and may be invalidated by subsequent discoveries, as well as resulting in too many 'overlaps' of genes belonging to more than one group. Among the most robust and long-lasting schemes devised so far are the standardized, 'phylogenetic' nomenclature systems presently in place for the plant kingdom and the *CYP*, *UGT*, *GST*, *SULT* and *ALDH* gene superfamilies. These systems satisfy the three criteria mentioned above. Their collective philosophy will permit the intelligent naming of, for example, a horse gene freshly cloned in 2008, reflecting homology and alignment studies. This assumes, of course, that rigorous attention has been paid to the naming of genes of other species in the meantime.

The apparent lack of concern of many people about the urgency of standardized gene nomenclature is unsettling. We do not have time to remain confused and uncommitted as to which direction we might take, with thousands of genes soon to be deposited in our databases. Either we prepare as quickly and efficiently as we can, or we remain in a muddle, arguing amongst ourselves, while disaster strikes. The power of an effective nomenclature system should not be underestimated, nor should the need for adequate resources to establish and maintain it. □

Julia White[1], Lois Maltais[2] & Daniel Nebert[3]
[1]The Galton Laboratory, University of London, UK. e-mail: julia@galton.ucl.ac.uk [2]The Jackson Laboratory, Bar Harbor, Maine, USA. [3]Center for Environmental Genetics, University of Cincinnati, Ohio, USA.