A piece of cake

Howy Jacobs

In the past year or so, I have devoted a lot of time to compiling manuscripts that incorporate outputs from genome-wide screens or profiles, which has required a lot of manual inspection of database entries. While those who work in computational biology may groan and advise me to let data-mining tools do the work—and do it more accurately, comprehensively, and objectively—I feel I have learned a lot, despite the excessive time invested. Most importantly, I have been able to absorb, digest, and condense the data into concepts that I could both understand myself and communicate to others, at least in a manner sufficient to satisfy peer-reviewers.

Most of us in molecular biology use the excellent resources of NCBI (the US National Center for Biotechnology Information) on a daily basis, courtesy of the US taxpayer, and partner organizations. Despite inevitable duplication and the constraints of a standardized annotation system, the information available from NCBI is generally clear and user-friendly. However, my overall experience of other publicly available databases of genomic, molecular, expression, and phenotypic information is less positive. Each of these resources tends to have its own virtues and idiosyncrasies, often beloved by the community at which they are primarily aimed, despite obvious limitations.

But many repositories seem to have some major, closely related deficiencies. The information they contain is typically over-inclusive and ‘flat’. In other words, there is little selection of what is really important or fundamental to each entry, with contradictory, outdated, anecdotal, and sometimes largely irrelevant data commonly placed alongside recent, systematic or experimentally rigorous findings. The naïve reader is left bewildered and befuddled by a mass of uninfused information, and making sense of it often requires one to find and evaluate the primary literature.

The simplest questions—what actually is this gene, protein, pathway, or structure, and what is its function—can take a lot of effort to answer. In many ways, this negates the very idea of having electronic compendia that we can rely upon. Consider the following example, admittedly a bit exaggerated, but which illustrates the problems encountered in trying to understand biological database entries.

Phenotypic information: Usually circular, commonly segmented with sixfold or eightfold rotational symmetry. Linear, triple-helical, coiled coil, and more complex morphs also observed.

Pathology: Strong association with type II diabetes, obesity, and other manifestations of metabolic syndrome. May predispose to cardiovascular disease. In extreme cases, patients may present with unexplained vomiting.

Composition: Variable stoichiometry: approx. 10% denatured protein, 65% carbohydrates, 25% fats (usually saturated), low amounts of vitamins and minerals. Traces of insect and fungal DNA reported. Core subunits often surrounded by a protective, highly glycosylated capsid containing neuro- attractants or other components mediating rapid phagocytosis.

Orthologs: Bos taurus—cheesecake, Gallus gallus—meringue, Drosophila melanogaster—fly cemetery, Pan troglodytes—stuffed monkey [1].

Paralogs: Homo sapiens—beefcake, Danio rerio—fishcake, Xenopus laevis—toad in the hole (paralogy disputed), Oryza sativa var. condoleezza (no supporting evidence)—yellowcake.

Interactions: Experimentally verified: B-52s, cup, wedding, urinal, walk, Motorcade of Generosity
Predicted: Marie Antoinette [2]; authenticity questioned by many authors.

Properties: Under unspecified conditions can be simultaneously had and eaten (untraceable author statement), strongly enhanced by icing on it, rapidly sold when hot (personal communication to Flanbase), and very bad when taken.

Related entries in national databases: Pound—UK, Viennoserie—Austria, Kake—Finland [3].

References:

A major problem with our organism-, organelle- and topic-specific databases is that they are mostly maintained on a shoestring budget by a few dedicated people. But they are every bit as vital to the quality of our work as core facilities offering state-of-the-art hardware for imaging, sequencing, or proteomics. To distill the literature and the mass of available data into an accessible but authoritative format, gene by gene, pathway by pathway, requires scholarship and broad expertise. It needs the judgement of someone sufficiently knowledgeable and impartial to select the key information and bury the mass of peripheral but still potentially meaningful findings at a deeper level. It cannot be done by a machine, a computer programmer, or even a world expert in a narrow area. It needs many of the same skills as a professional editor. Above all, it needs secure funding and an international framework. If we can achieve all that, interrogating the literature about a novel gene that came out of a screen will hopefully one day be a piece of cake.