By Jeffrey M. Perkel

LAB TOOLS

Why You Should Be Annotating

Scientists who rely on accurate gene predictions should share in the burden of creating them

Your local convenience store probably has a dish filled with pennies near the checkout. If your order costs $1.01 and you don't have a penny, you take one. The next time you're in, if you get change, you're expected to leave a penny. Unfortunately, when it comes to annotating sequence databases, it seems most researchers are the type to take a penny, but not give one back.

With the click of a mouse, scientists gain free access to enormously expensive and annotated sequence databases, the product of teams of researchers and informaticians. Yet when users notice errors in annotation - gene models that don't match their own data, for instance - they generally keep the knowledge to themselves. "There's a certain amount of apathy," says WormBase developer Lincoln Stein. "People realize a gene model is incorrect but they don't report it."

They don't alert database curators to new gene models much, either. WormBase, which serves tens of thousands of visitors per month, receives "a small but steady stream of feedback, on the order of a few per week," says Stein. PlantGDB, with about 2,800 unique visits per month, has received about 200 annotation submissions in the past year.

Clearly, more help is needed. With some 800 genomes in the sequencing pipeline, the resources simply aren't there to give every organism its own annotation team. The Berkeley Drosophila Genome Project (BDGP) employed a bicoastal group of 10 to annotate the fly's 176-megabase genome three separate times. Armed with extensive genetic and experimental data, including some 255,000 expressed sequence tags (ESTs) and more than 9,000 full-length cDNAs, annotators spent a full year on each annotation cycle. And they aren't finished: Annotation version 4.3, released January 30, includes about 100 new and updated gene models.

Suzanna Lewis, former head of informatics at BDGP, puts the group's annotation budget for 2001 at around $500,000. Lacking such resources, most genomes will end up in automated annotation pipelines instead. Yet the resulting predictions, unfortunately, are

largely unreliable.

Computational gene-finding algorithms combine ab initio predictions, homology data, and experimental data to create gene-model predictions. Last year, in an attempt to gauge how well software performs at that task, Roderic Guigo Serra, professor of bioinformatics at the University of Pompeu Fabra in Barcelona, held a human genome annotation competition called EGASP. The 18 competing algorithms did reasonably well in predicting exon boundaries relative to human curators, says Guigo, but even the best programs could correctly string those exons together into transcripts only about 40% of the time. The algorithms, "are unable to reproduce what the human being is able to do when trying to put these exons into transcripts," says Guigo, who expects to publish the EGASP findings later this year.

One way to improve gene annotation is to set aside genome funding to shore up the data available to computational pipelines. Researchers can use high-density tiling arrays, for instance, to identify transcribed sequences across the genome; 5' RACE to pinpoint transcription start sites; and cDNA libraries to define intron-exon boundaries. "Every person I talk to from a genome sequencing center says, please, when you [request money to] sequence a genome, also support some funding for ESTs, because it's no good doing the genome without doing some EST or cDNA sequencing as well," says Lewis.

Still, accuracy will be maximized only if the community chips in. Model organism databases generally provide mechanisms to add new gene models, or correct errant ones, either via E-mail or Web-based forms. Gene-annotation wikis could fill the same role for orphan organisms that lack dedicated databases (Nature, 439:534, 2006).

Peter Good, program director at the National Human Genome Research Institute, tells of a colleague who actually offered T-shirts to people who submitted gene annotations to a yeast database. "Very few people took him up on it," he says.

Sounds like the dish could use a few more pennies.

jperkel@the-scientist.com

Comment on this article