Annotation competition spurs *Drosophila* sequencing efforts

[HEIDELBERG] Bioinformaticists have used a competition between rival research teams to identify the computational modelling tools that can most accurately locate and describe the genes hidden within a stretch of sequenced genome.

The results of the competition were assessed last week at the seventh international conference on Intelligent Systems for Molecular Biology (ISMB) in Heidelberg, Germany. The idea had been to stimulate the bioinformatics community to improve the predictive power of its software tools.

The immediate goal was to help identify the best tools for annotating the *Drosophila* genome, the sequence of which is due to be completed by the end of the year.

The idea of a competition came from Martin Reese, a postdoctoral fellow with the Berkeley *Drosophila* Genome Project (BDGP). He modelled it on the Critical Assessment of Techniques for Protein Structure Prediction (CASP) project, a biennial competition which asks computational modellers to predict the three-dimensional structure of proteins from DNA sequences.

Every other year, CASP competition organizers select proteins that are sequenced and whose structures are confidently expected to have been determined by X-ray crystallography by the time the competition ends (see *Nature Structural Biology* **6,** 108; 1999).

In recognition of CASP, Reese has named his own project GASP — Genome Annotation Assessment Project. He says he wanted to "see how well our science community could work together". But, unlike CASP, he says, the annotation project cannot be considered a real competition as there is no absolute 'right' answer, so no winners will be announced.

Earlier this year, the BDGP reached an agreement with the genomics company Celera to collaborate on sequencing and annotating the *Drosophila* genome, to which Celera is applying its 'shotgun' approach (see *Nature* **393,** 296; 1998). Celera is a joint venture between the Perkin Elmer Corporation and The Institute for Genomic Research, directed by Craig Venter.

The BDGP is taking a slower but more thorough approach to the sequencing, and has already confirmed 25 megabases of the approximately 140-megabase genome. The Berkeley team will help Celera piece together its jumble of data. Celera got off to a late start because the delivery of 3700 DNA Analyzers produced by Applied Biosystems, part of Perkin Elmer, was delayed. Celera says it will begin making raw data public in October.

WALTER SMALLING

Warming up: Celera executives Mark Adams (right) and Hamilton Smith in the company's sequencing lab. High-quality software is vital for accurately sifting genes out of raw sequence data.

In May, Reese and BDGP project leader Gerald Rubin chose three megabases of their own unpublished sequence for the competition. Their own annotation, identifying and locating 220 genes in the stretch, was used as a reference. The competition was judged by an international panel of geneticists and bioinformaticists.

Rubin says he was "very pleasantly surprised" that 12 research teams around the world took up the challenge. Participants, who were given only six weeks to respond, say that they were equally happy to have the opportunity to benchmark their software tools against those of other groups.

One participant, Terry Gaasterland, a bioinformaticist at Rockefeller University in New York, says that the competition was "an amazing motivation for our group," which had not used its software tools to approach such a complex organism. "We wanted to win, even knowing there is no real winner," she says. "We would certainly do it again."

One unexpected result, says Reese, is a consensus that the Berkeley group may have underestimated the number of genes in the stretch. Most participants predicted 20 or 30 more genes than the group had found.

Over the next six months, biologists at Berkeley will check the authenticity of all the predicted genes by seeing if the relevant sequences really do direct the synthesis of proteins in laboratory experiments.

"Annotation helps drive biological experiments," says Rubin. And bioinformaticists find the feedback from experiments — "the reality check" — very encouraging, he says.

Equally importantly, the information about the quality of the software products being developed by different groups around the world will aid efforts to annotate the *Drosophila* genome by highlighting the best tools. "The competition has been an excellent warm-up for the larger task of annotating the whole genome, and setting standards for the completeness and quality of the annotation," says Mark Adams, Celera's vice-president for genome programmes.

For Celera and many BDGP scientists, the *Drosophila* genome is only a prelude to the human genome, whose full sequence Celera, engaged in sharp competition with publicly funded sequencing efforts, expects to complete by the end of 2001. "Our work on *Drosophila* will provide feedback for the Human Genome Project on how well shotgun sequencing works and on how best to annotate," says Rubin.

The annotation competition will probably be repeated at a future ISMB conference. Reflecting the growing importance of informatics in biology, the ISMB is trying to cope with an almost geometric increase in attendance at meetings in recent years.

Thomas Lengauer, an institute director at the German Research Centre for Information Technology in Bonn, is delighted with the growth of interest. But he hopes that weight of numbers will not force organizers to hold more informal meetings where presentations are not refereed in advance. At Heidelberg, 35 papers were selected for presentation from the 150 submitted.

The ISMB will be sponsored in future by a new umbrella organization, the International Society for Computational Biology. Russ Altman, a bioinformaticist from Stanford University in California, was last week chosen as the society's president-elect, and Lengauer its vice-president-elect.

Alison Abbott