Open annotation offers a democratic solution to genome sequencing

Sir — Jean-Michel Claverie[1] writes in Correspondence about the problems of annotating the whole human genome sequence, given that a draft form will be available in a few months. While we agree with many of his points, we disagree with what he says about the lack of bioinformatics capacity to provide a useful basic analysis. The Sanger laboratories, with the European Molecular Biology Laboratory's European Bioinformatics Institute, have been developing an automatic analysis system for some months; the results of the first full release of Ensembl can be seen at http://www.ensembl.org/. The system now tracks the daily output of human genomic sequence in real time. It is based on confirming *ab initio* predictions by homology and providing functional annotation via Pfam[2]. So far 17,045 gene fragments are annotated from the 1,405,539,258 bases processed.

We agree with Claverie about the limitations of any automatic analysis system, having ourselves worked on the semi-manual analysis of the human chromosome 22 sequence. However, a large subset of genes can already be predicted accurately, which will be very useful as a way into this huge volume of data. A key aspect of the system is its ability to keep track of genes despite revisions to the sequence. This will be important as the genome is completely sequenced over the next couple of years. Ensembl accession numbers assigned to genes are permanent identifiers that will refer to the same genes throughout this process.

How can we go beyond this baseline automatic annotation? Claverie points out the chaos that would result from duplicated annotation efforts, each with different standards and different ways of presenting the data. He is also correct in arguing that no single collaborative group will be capable of annotating the entire genome consistently and to high quality. One way to deal with this is to have a monolithic single entity that invests 300 person-years into annotating the genome. A better one is 'open annotation', where the annotation required is distributed across a highly motivated community of biologists.

We believe that many of the problems with open annotation are technical ones, which can be and are being addressed. The web allows different data sources to be readily crosslinked, but different websites have different formats and interfaces. An alternative, particularly appropriate for sequence data, is for a browser to merge annotation from multiple data sources on top of a baseline coordinate system to provide the user with a single annotation view. Lincoln Stein and colleagues are developing such a system (DAS) based on XML (see http://stein.cshl.org/das/). All that is then required for any centre to contribute annotation of all or part of the genome is to synchronize its coordinate system with its baseline server. Maintaining the coordinate system across a changing genome does require substantial resources, but keeping in synchronization with this need not. Ensembl is an open-source project and will provide both a common object framework for annotation as well as the synchronization tools needed for anyone to set up to serve annotation for all to see and use.

The power of open-source software is well recognized[3], although it could be feared that open annotation will swamp biologists with alternative contradictory views of the sequence. We are more optimistic. Browsers will allow biologists to select only the data sources they wish to view. Just as some websites become popular, word of useful annotation will spread quickly, since selecting it will be as easy as bookmarking a new website. Software development has been democratized by open-source projects such as Linux, which have allowed everyone the opportunity to contribute. Open annotation provides the same opportunity for genomes, and so should speed our collective decoding of genetics without centralized annotation centres or commercial monopolies.

Tim Hubbard*, Ewan Birney†
**Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK*
†EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK

1. Claverie, J.-M. *Nature* **403,** 12 (2000).
2. Bateman, A. *et al. Nucleic Acids Res.* **28,** 263–266 (2000).
3. *Nature* **403,** 229 (2000).

Affirmative action won't solve sex discrimination

Sir — Natasha Loder's article and the accompanying cartoon on gender discrimination (*Nature* **402**, 337; 1999) shed little light on this vexed issue. Some people have no doubt that discrimination, sexual and otherwise, does exist in academic institutions, as it does in most other human endeavours. Others consider this merely a reasonable working hypothesis requiring clear evidence, the type of evidence that the European Technology Assessment Network report attempts to provide.

The excerpt from the report presented in the article, concerning the low proportion of women in national scientific academies, is unconvincing to anyone, male or female. It is clear that there are fewer women in the upper echelons of academic research, but there are many possible reasons. The most parsimonious of these is that the long climb up the academic ladder takes a few decades, and the present demographics in national scientific academies reflect newly trained scientists emerging 20 or 30 years ago.

If we accept that gender-based discrimination is wrong, we should at least try to examine the problem more rigorously before suggesting sexually discriminatory policies aimed at ensuring a gender balance on public bodies. If, indeed, time-lags are responsible for the gender disparities, they will disappear in due course, independently of changes in hiring and funding practices. Discrimination is the problem, not the solution.

G. A. Lozano
Department of Biological Sciences, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

Why Roche deserves the disputed *Taq* patent

Sir — You report[1] the legal decision in the long-running Roche–Promega dispute, that the US patent on full-length *Taq* DNA polymerase is invalid. The patent (4,889,818) had been awarded to Cetus Corporation and subsequently bought by Roche. This decision does not affect the validity or enforceability of Roche's foundational patents on polymerase chain reaction (PCR) and other related patents, including those for using any thermostable enzyme, including native *Taq*, for PCR.

I believe that the court's decision was wrong and unfair to Cetus scientists David Gelfand and Susanne Stoffel, in that it did not distinguish their invention from the work of the Gorodetskii[2] and Trela[3] groups. Cetus scientists including Gelfand and Stoffel were the first[4] to isolate and clone the full-length (molecular mass 94,000; 94K) *Taq* DNA polymerase. The earlier groups repeatedly published their isolation of *Taq* fragments (60K–70K), undoubtedly the result of proteolytic degradation, under the mistaken impression that it was the complete enzyme.

Instead of concentrating on the validity of the Cetus invention, Promega's case was based on misrepresenting the raw experimental data of the scientists and their good-faith interpretations of it. By this stratagem, Promega is trying to rewrite history by asserting that Cetus had data indicating that the earlier groups isolated a 94K