

Do we need a huge new centre to annotate the human genome?

Sir— Now that a human chromosome has been sequenced¹, and a rough draft of the complete genome promised for the spring², one obvious requirement is to annotate the genome; that is, to precisely locate all the genes and assign them to their protein products. There are sure to be dozens of applications for “human genome annotation centres”, all emphasizing the strategic nature of this endeavour and the huge expected benefits for national biomedical industries. How can we avoid a planetary-scale duplication of effort caused by the creation of multiple human genome databases, all using different protocols, software and format?

Paradoxically, the high number of candidate centres arises because we do not yet have the bioinformatic capacity to annotate 3 gigabases of human genome sequence accurately. Although the best gene-finding programs can detect protein-coding exons with 80% accuracy³, performances on identifying 5' and 3' untranslated regions, core promoters or other regulatory sequences are close to zero. Moreover, alternative transcript forms (probably involving more than 30% of genes) are not addressed by any program, and non-protein-coding genes (for example *XIST*, *IPW* or *NTT*) are totally transparent to non-similarity-based detection programs³.

A recent complementary-DNA-based survey of *Caenorhabditis elegans* (nematode worm) annotations reported by Jean Thierry-Mieg⁴ shows that only 50% of proteins are correctly predicted, the rest exhibiting various degrees of annotation errors (for example, incorrect, added or missed exons; fused genes; or being entirely wrong). The *C. elegans* genome is more compact, much smaller and much less repetitive than the human. An impressive gathering of leading computer biologists has therefore recommended the sequencing of more expressed sequence tags (ESTs) and cDNAs as the best way to annotate the human genome⁴.

In the absence of accurate automated methods, the level we can hope to achieve will rely on a fairly trivial protocol, mostly based on similarity searches, which could be set up by any competent bioinformatics laboratory within weeks. The cost of the required computing power is well under \$300,000, if a PC cluster solution is retained. For instance, a cluster of 125 Pentium 3 (see “gigablast” at <http://igs-server.cnrs-mrs.fr>) would take one day to compare all human EST contigs to the whole human genome sequence, and the rest of the week to compare it with all known proteins.

However, the simplicity of protocols like this one is also a recipe for disaster. At each step, a variety of programs, arbitrary parameters and thresholds can be chosen. Multiple, independent genomic annotation efforts would thus result in highly redundant human genome annotation databases.

Going further than the above automated annotation stages will require a lot more money and a bigger organization. For example, detailed visual inspection of each gene will be needed to reduce the 50% error rate significantly. I estimate the workload at one gene each day per annotator: 300 person-years in total. Starting from a unique, single-pass, automatically annotated human genome sequence, such work could be split among 30 international groups of five people each. This would complete the work in two years.

What about the “strategic value” of an annotated human genome sequence? Does it justify launching multiple efforts?

I think not, for two main reasons. First, most pharmaceutical companies seriously involved in genomics have already identified many targets using cDNA and EST data from private sources. It is unlikely that many new genes of economic value will be revealed by the systematic annotation of the human genome (especially if public EST and cDNA data are central in the protocol).

Second, increasing our knowledge of human genomic sequence at a large scale is sure to generate significant improvements in the next generation of gene-finding software. It would then be ridiculous to have embarked on a detailed annotation effort using obsolete methods.

Jean-Michel Claverie

*Structural & Genetic Information Laboratory,
CNRS-UMR 1889, Marseille, France*

1. Dunham, J. *et al. Nature* **402**, 489–495 (1999).
2. Wadman, M. *Nature* **398**, 177 (1999).
3. Claverie, J.-M. *Hum. Mol. Genet.* **8**, 1821–1832 (1999).
4. Cold Spring Harbor First Workshop on Computational Biology: Bridging the Gap between Sequence and Function (7–9 September 1999).

How nature itself uses genetic modification

Sir— Mae Wan Ho¹ states that genetic engineering is fundamentally different from conventional plant breeding or wide crosses to produce novel crops; thus she concludes that genetically modified (GM) plants need unique tests on their safety.

Any new crop should have its composition of major and minor constituents investigated to establish substantial equivalence, and its novel trait independently tested, as has been the case for the present range of GM plants. However, Ho states that special tests are required because “genetic engineering enables exotic genes... [to be]

combined in novel constructs, often with viral promoters to make genes overexpress continuously. The constructs are inserted into genomes by transformation techniques that cannot control where the genes go, resulting in a range of unpredictable positional effects and rearrangements”².

But this differs little, if at all, from Ho's description of normal genome behaviour in her book², which states: “Genome organisation is infinitely variable [and contains] transposons that can excise and reinsert elements in different locations in the genome” and “up to 20% of some genomes may contain reverse transcripts. These processes destabilise genes and genomes, move genes around, mutate, rearrange, recombine, replicate sequences...”. What could be more exotic than a mutant protein?

Some estimates suggest that the genome of some cereals may contain up to 50% retrotransposons; transposons contain end regions that are hot spots for recombination using transposase. The plant genome contains very large numbers of strong promoters that direct expression as strongly as any viral promoter such as the cauliflower mosaic virus 35S promoter. These promoters are used experimentally; many of them are constitutively and continuously expressed because they control the expression of housekeeping genes. We know the consequences of events via random GM insertion from libraries constructed from T-DNA insertion. However, Ho and other critics of GM technology neglect the extent to which selection is made among many GM transformants, as is the case with conventional plant breeding among siblings. Lethal insertions are self-selecting; potentially innocuous insertions are detected by substantial equivalence.

It is important to recognize that all the food we eat has been (and is) continuously genetically engineered by natural phenomena in ways that do not differ in any fundamental way from the current GM technology. Natural genetic modification of wheat³ and rice, for example, enabled the breeding of dwarf crops, used to feed many millions in the ‘Green Revolution’.

To criticize experimental GM technology while accepting the benefits of natural GM implicit in ‘the fluid genome’² hints at a not uncommon attitude that sees synthetic pesticides as morally reprehensible but natural pesticides as good. Is this really any different from notions of Original Sin?

Anthony Trewavas

*Institute of Cell and Molecular Biology,
University of Edinburgh, Edinburgh EH9 3JH, UK*

Christopher Leaver

*Department of Plant Sciences, University of Oxford,
South Parks Road, Oxford OX1 3RB, UK*

1. Ho, M.-W. *Nature* **402**, 575 (1999).
2. Ho, M.-W. *Genetic Engineering: Dream or Nightmare?* (Gateway, Bath, 1998).
3. Peng, J. *et al. Nature* **400**, 256–261 (1999).