

Showcase on Research

The *bête noire* of the Malaria Genome Project: A Personal View

Robert Huestis

Department of Microbiology and the Victorian Bioinformatics Consortium, Monash University, Clayton, VIC 3800

Annotation is the *bête noire* of sequencing projects (1). With the publication of the draft sequence of the human genome, now begins the struggle to work out the annotation. "The sequence is not an end, but a beginning. ...As yet, the detection of genes, even in finished sequences, is still in its infancy" (2). At best, annotations vary by 30% (3). In the latest review 35,000, 65,000, 80,000 and 90,000 genes are predicted from the same sequence by different investigators. Laboratory testing of a random sample is proposed to see who is closest to the real number (4). Truly, annotation is the "black beast" of sequencing. As with the Human Genome Project, annotation has turned out to be the *bête noire* of the Malaria Genome Project. Why is this?

A gene without introns is relatively easy to annotate. It is found by deduction. The region is compared to the statistical model of a coding region (5, 6); the first ATG in the open reading frame is the initiation codon; the stop at the end of the open reading frame is the end of the gene. *Plasmodium falciparum* was initially thought to be little more advanced in the evolutionary sense than a bacterium, and to contain few introns, and then typically only one or two per gene where they did exist. The emerging picture of *P. falciparum* is quite different. Most genes have introns. I have predicted one gene in chromosome 14 with 23 exons (7). The maximum number is unknown.

From this overview we can see that the Malaria Genome Project was initially planned for a certain number of years on the assumption that annotation could be successfully completed with a good gene finding computer program. GLIMMER analysed chromosome 2 in 50 minutes (8). Then complications began. The annotators of chromosome 2 proposed some additional genes in chromosome 3 that had not been read in the initial annotation (9, 10). As with all but experimentally verified annotation, these proposals were only predictions. In the lab some predicted

genes produce cDNAs; some do not (11). There is nothing unusual in this result.

In science there is no guarantee that anyone will ever discover anything. A project outline is not a promise but a plan. The Greek word for "I promise" is literally "I plan". However, it was assumed that some computer algorithm could magically find the genes. It is assumed today that a computer algorithm can find the genes, if only that algorithm can be trained with a large enough and accurate enough training set.

An airline pilot manages risks according to the worst case. Suppose that in the worst case this assumption is not true. Suppose there is a "*bête noire*" lurking in the shadows. Suppose that the more introns we analyse, the more variation we will find; suppose the more variation in our training set, the more confused our algorithms become; suppose, because computers operate deductively (12), comparing the observed to the expected, there is a barrier of ~80% of genes read, a barrier the algorithms cannot penetrate. Suppose...

In this worst case there will be an overrun in terms of both time and cost. Is there anything wrong with such an outcome? No. In science no investigator can guarantee to find Sarendip or Shangri-La. We do not promise the funding agencies any particular outcome; we plan.

The present state of the Malaria Genome Sequencing Project is that the majority of the sequence is completed, but the majority of the genes are not yet annotated. This month the sequencing centres will begin a coordinated effort to annotate all of the completed sequence with the best software available, planning to make the results public at the beginning of 2002. This plan depends upon the presumed ability of the best computer algorithms together with the best available training data to be able to find the genes. I do not expect the difficult genes to be found. I believe the computer has an upper limit of ~80% of the genes.

I believe that 2002 will mark the beginning of the annotation of *Plasmodium falciparum*, not the end. I believe the current state of the Malaria Genome Sequencing Project is a condition of hope which may not be realised. We may complete the sequencing only to discover we have not trained even one analyst find the genes. Today, science is so thoroughly allied with the computer that manual methods are not in tune with society's thinking. Secret messages, of which gene messages are one kind, are uncovered by the use of precise logic combined with a touch of serendipity (12, 13).

Two objections are raised against manual annotation. First, is manual annotation objective? Can the results be replicated? This question has been discussed comprehensively and settled in the cryptology literature. Two analysts using standard techniques of cryptanalysis, suitably trained, must uncover the same message (12, 14, 15). While in cryptanalysis there is only one legitimate message, in gene-finding the situation is sometimes complicated by alternative splicing. Or the gene may be so unusual that only the parasite knows where to splice. Yet mainly, the cryptanalytic dictum holds as follows. Any analyst suitably skilled must be able to find all genes for which there is sufficient evidence. In cryptanalysis a message must be of a certain length to permit a unique solution (15); in gene analysis there must be sufficient evidence to assure a unique solution.

Second, it is argued that there is not time for manual analysis of the estimated 6000 to 7000 malaria genes. If we would invest one year training one analyst, on the average, for each of the 14 chromosomes, the entire malaria genome could be accurately annotated within that year. Manual annotation requires "perseverance, careful methods of analysis, intuition and luck in the order named" (12, 16, 17), and takes, by my estimate, three to four weeks per Mb at the hands of a skilled and experienced analyst.

Showcase on Research

The bête noire of the Malaria Genome Project: A personal View (contin.)

Let us train analysts to find genes now, not only in malaria but in every sector of genetic research. Let us factor into the planning of the Malaria Genome Project a time allotment for human annotation of the genome. Let us avoid a time- and cost- overrun.

Acknowledgements

This research was performed with assistance from the Australian National Health and Medical Research Council and the Queensland Government, the Victorian Bioinformatics Consortium and the Wellcome Trust as part of the PlasmoDB project.

Editor's note: In view of the special nature of this article and the literary references and historical allusions, citation of references does not follow the standard format for the Australian Biochemist.

References

1. Marocco, W.T. (1967) *Musica ficta* is the bête noire of transcribers of fourteenth and fifteenth century music. *Italian Secular Music. Polyphonic Music of the Fourteenth Century* (vol.VI), Éditions de L'Oiseau-Lyre. Les Remparts, Monaco
2. Sulston, J. (2001) *Biochem. Soc. Trans.* **29**, 27-31 [Sir Frederick Gowland Hopkins Memorial Lecture delivered at the University of Sussex on 20 December 2000]
3. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860 - 921
4. Kintisch, E. (2001) So what's the score? *New Scientist* **16**, 2290
5. Huestis, R., and Saul, A. (2001) An algorithm to predict 3' intron splice sites in *Plasmodium falciparum* genomic sequences. *Mol. Biochem. Parasitol.* **112**, 71-77
6. Saul, A., and Battistutta, D. (1988) Codon usage in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **27**, 35-42
7. Institute for Genomics Research contig c14m8, t42w + t43w + t44w: 151816..151822, 151966..152084, 152197..152329, 152472..152613, 152708..152786, 152974..153147, 153320..153369, 153517..153575, 153682..153891, 153993..154166, 154240..154343, 154493..154708, 154824..154894, 155001..155071, 155170..155267, 155366..155442, 155532..155610, 155701..155886, 155988..156045, 156160..156251, 156436..156489, 156658..156748, 156844..156878
8. Gardner, M., Tettelin, H., Carucci, D., Cummings, L., Aravind, E., and Hoffman, S. et al. (1998) Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126-1132
9. Pertea, M., Salzberg, S.L., and Gardner, M.J. (2000) Bioinformatics: Finding genes in *Plasmodium falciparum*. *Nature* **404**, 34
10. Lawson, D., Bowman, S., and Barrell, B. (2000) Bioinformatics: Finding genes in *Plasmodium falciparum*, reply. *Nature* **404**, 34-35
11. Huestis, R., and Fischer, K. (2001) Prediction of many new exons and introns in *Plasmodium falciparum* chromosome 2. *Mol. Biochem. Parasitol.* (in press)
12. Kahn, D. (1996) *The Codebreakers* (2nd ed.), Weidenfeld and Nicolson, London
13. Kippenhahn, R. (1999) *Code Breaking: A History and Exploration*. Trans. E. Osers. Overlook Press, Woodstock
14. Friedman, W. F., and Friedman, E. (1957) *The Shakespearian Ciphers Examined*. Cambridge University Press
15. Shannon, C.E. (1949) Communication theory of secrecy systems. *Bell Syst. Tech. J.* **28**, 657-715
16. Hitt, P. (1916) *Manual for Solution of Military Ciphers*. Signals Intelligence Service, Government Printing Office, Washington DC
17. Callimahos, L., and Friedman, W. (no date - classified) *Military Cryptanalytics*. Reprinted as Cryptographic Series #42, 1956 (Laguna Hills CA, Aegean Park Press, 1985, p. 18, footnote). Revised and annotated version of W. Friedman, *Military Cryptanalysis* (classified, n.d., ca. 1937). Reviewed in Callimahos, "Cryptology: cryptanalysis", *Encyclopedia Britannica*, 15th ed., 1977

