EvoBio 2012, M. Giacobini, *et al. eds.*, LNCS 7246, 245-249, Malaga, 11-13 Apr, Springer

In Silico Infection of the Human Genome

W. B. Langdon and M. J. Arno

Dept. of Computer Science, University College, London, Gower Street, WC1E 6BT, UK

Abstract. The human genetic sequence database contains DNA sequences very like those of mycoplasma bacteria. It appears such bacteria infect not only molecular Biology laboratories but their genes were picked up from contaminated samples and inserted into GenBank as if they were homo sapiens. At least one mouldy EST (Expressed Sequence Tag) has transferred from online public databases on the Internet to commercial tools (Affymetrix HG-U133 plus 2.0 microarrays). We report a second example (DA466599) and suggest there is a need to clean up genomic databases but fear current tools will be inadequate to catch genes which have jumped the silicon barrier.

Keywords Bioinformatics, data cleansing, bit rot, *in silico* biology, meme, Blast, phishing, jumping information genes, *in silico* evolution

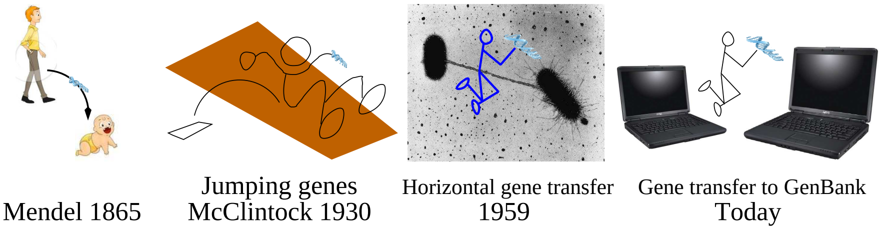

Fig. 1. Tetratych showing: 1865 Mendel's [1] discovery of the essential digital nature of inheritance; 1930 Barbara McClintock's [2] discovery of transposons in Maize whereby genes move not only from parent to child but also along chromosomes; 1959 Micrograph of genetic transfer along a pilus linking two bacteria (Akiba and Ochia discovered the first interspecies gene transfer [3]); mycoplasma bacteria genes are transferred between computers, including into the reference human genome DNA sequence held by GenBank [4].

1 Introduction

Figure 1 shows how our understanding of genetics has changed over the last 150 years. In each frame the small blue double helix is used to illustrate the movement of genes. In 1865 Mendel showed that genes are discrete units. Originally it was thought that genes were inherited only from parents, however it is now known that genes can be transferred horizontally. Firstly in 1930 McClintock showed genes could move to new positions in chromosomes of the same species. Later it was found that jumping genes could be transferred between species [3]. Indeed today lateral gene transfer mediated by viral agents is thought to be common [5].

What so far has been little recognised is that jumping genes have escaped biology and now roam our computer systems.

2 The Human Genome

Ensuring databases are both up to date and contain only correct data is a huge software engineering problem. Even as the human genome was first published the associated problems of data cleansing Bioinformatics sequence data were being discussed [6; 7] but it appears only technical problems where considered.

We discovered that GenBank, the definitive publicly accessible database holding the human DNA sequence, has been corrupted in a surprising way. It contains the DNA sequence of a bacteria [4].

Section 3 describes how we recently discovered a second sequence which is probably not human in the human genome [8]. Here we extend RN/11/14 [8].

It appears that not only has the human DNA sequence been "completely sequenced" [6] but in the process other living organisms commonly found in molecular biology laboratories have infected not just the physical samples but also the virtual *in silico* Bioinformatics environment. By unwittingly using a technique reminiscent of computer hacking, a bacteria gene has succeeded in not just moving within its own genome [2] nor only jumping horizontally and crossing the species barrier [3] but has crossed the silicon barrier between life and data and succeeded in reproducing itself across very diverse information based media. Given the highly interconnected nature of genomic research, technology and medicine and the low priority so far attached to the problem, it is unlikely current data warehouse cleansing techniques will be able to eradicate this and potentially other silicon jumping genes.

3 Computational *in silico* Experiment

Using Blast [9] at the European Bioinformatics Institute with their default settings, we searched for the anomalous HG-U133 +2 gene sequence (GenBank AF241217, probeset 1570561_at, which we reported in [4]). This gave a list of DNA sequences which partially match published DNA sequences. The list is ordered by blastn so that the best matches are at the top. Only the top 50 fuzzy matches are included. As expected the first match is the query sequence itself (EM_HTG:AF241217). Despite [4] having been published more than a year ago, EM_HTG:AF241217 is still described as "Homo sapiens". All the others are mycoplasma, except the 34^{th} in the list, DA466599, which EBI says is human. (EBI gives one reference for DA466599: [10]. DA466599 was uploaded to the DNA Data Bank of Japan 2 years after the HG-U133 +2 was launched). However we suggest that DA466599 may not be a human DNA sequences but is another example of physical contamination leading to virtual infection of the public data.

We ran a second EBI blastn query (again using the NCBI em_rel database). This time looking for DNA sequences that match DA466599. The results for DA466599 are similar to those for AF241217 and so support the view that DNA sequence DA466599 is not human but instead is also a contamination. Again the best 50 matches were reported. Of course the first one is DA466599 itself. All the other matches returned by blastn are for various species of mycoplasma.

4 No *In Silico* Evolution?

Notice what these mycoplasma genes have done. Not only, despite rigorous hygiene standards, do they routinely spread themselves through microbiological laboratories [11] but now at least two have got themselves copied into GenBank and one has spread from there into an Affymetrix GeneChip design. How is this different from any other case where a gene has been sequenced? Fundamentally it is the same. But notice, even though we can post hoc guess a mechanism, it is as if the gene had acted to spread itself. In Biology, gene DNA sequences are acted upon by many mechanisms that copy them but we still adopt the short hand of saying the gene has spread itself [12].

It is difficult to know the number of copies of the human genome. However if we ignore the small number of mirror sites and assume everyone downloads sequences from GenBank directly. This means the GenBank's Internet bandwidth limits the number of copies. Since each copy takes 2 hours, the maximum downloads per year is 4383. Although new versions of GenBank are released "every two months" GenBank is fairly stable and people may not need to be fully up to date, therefore we suggest each copy lasts about a year. This gives an estimate of the global population of the human genome DNA sequence of less than 4 000.

In biology none of the DNA copying mechanisms is perfect. This ensures inherited material is subject to variation. In our computer systems it is often assumed copies are perfect. Indeed we have seen no evidence, yet, of DNA sequences being corrupted once they have been captured by our databases. However changes are possible. Rosenthal [13] says "1.2 10^{-9} of the data written to CERN's storage was permanently corrupted within six months". Also error rates on transferring data across the Internet are never better than 10^{-12} [14] and wireless connections to portable devices are very much worse. Even in the best cases, operator error is always a hazard [14]. In other words, error rates in the best computer systems are much less than typical mutation rates but accidental changes are possible, particularly with portable laptop computers.

After reproduction and variation, the third requirement of evolution is selection. Although one might see human imposed differential selection on corrupted gene sequences, the most likely selection pressure would be simply aimed at removal of errors. Complete extermination would not lead to evolution. Partial erasure might serve. However given the small population size and hence low copy rate, very low mutation rates and absence of suitable fitness selection, the conditions for the evolution of these *in silico* genes are currently poor.

5 Discussion

It is well known that mycoplasma contamination is rife [11]. Many labs are routinely periodically sterilised to counter it. Miller *et al.* [11] said mycoplasma contamination has "potentially major consequences for the diagnosis and characterization of diseases using expression array technology." Even so, using RNAnet `http://bioinformatics.essex.ac.uk/users/wlangdon/rnanet/`, we previously estimated about 1% of published data in the Gene Expression Omnibus (GEO) database at NCBI (`www.ncbi.nlm.nih.gov/geo`) are contaminated [4].

One potential fortuitous side effect of the *in silico* spread of mycoplasma contamination is that the Affymetrix HG-U133 +2 1570561_at probeset might be used to indicate physical sample contamination. Thus probeset 1570561_at could be treated as a free additional quality control signal. If 1570561_at says there is significant expression of mycoplasma genes, then the sample is probably contaminated and the other gene expression levels given by the microarray are suspect.

Having found two suspect DNA sequences it seems likely the published "human genome" sequence contains more. Indeed contamination of all organism sequences seems possible [15]. With the explosive growth of genomic sequence data available via the Internet, including data from the 1000 genome project [16], it seems time to look again at genomic database quality.

References

1. Gregor Mendel. Experiments in plant hybridization. *Verhandlungen des naturforschenden Vereines in Brno*, (IV):3–47, 1865. Translated by William Bateson.
2. Barbara McClintock. A cytological and genetical study of triploid maize. *Genetics*, 14(2):180–222, 1929.
3. Tomoichiro Akiba, *et al.* On the mechanism of the development of multiple-drug-resistant clones of Shigella. *Japanese Journal of Microbiology*, 4:219–227, Apr 1960.
4. E. Aldecoa-Otalora, W. B. Langdon, P. Cunningham, M. J. Arno.Unexpected pre sence of mycoplasma probes on human microarrays. *BioTechniques*, 47(6):1013–16.
5. Lauren D. McDaniel, *et al.* High frequency of horizontal gene transfer in the oceans. *Science*, 330(6000):50, 1 October 2010.
6. Adam Felsenfeld, *et al.* Assessing the quality of the DNA sequence from the human genome project. *Genome Research*, 9:1–4, 1999.
7. Rolf Apweiler, *et al.* Technical comment to "Database verification studies of SWISS-PROT and GenBank" by Karp et al. *Bioinformatics*, 17(6):533–534, 2001.
8. W. B. Langdon and M. J. Arno. More mouldy data: Another mycoplasma gene jumps the silicon barrier into the human genome. *ArXiv e-prints*, 14 June 2011.
9. Stephen F. Altschul, *et al.* Gapped BLAST and PSI-BLAST a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, 1997.
10. Kouichi Kimura, *et al.* Diversification of transcriptional modulation: Large-scale identification and characterization of putative alternative promoters of human genes. *Genome Research*, 16:55–65, 2006.
11. Crispin J. Miller, *et al.* Mycoplasma infection significantly alters microarray gene expression profiles. *BioTechniques*, 35(4):812–814, October 2003.
12. R. Dawkins. *The Selfish Gene.* Oxford University Press, Oxford, 1976.
13. David Rosenthal. Keeping bits safe: How hard can it be? *Commun ACM*, 53:47–55.
14. Mark Handley. Why the internet only just works. *BT Technology Journal*, 24(3).
15. Mark S. Longo, *et al.* Abundant human DNA contamination identified in non-primate genome databases. *PLoS ONE*, 6(2):e16410, 02 2011.
16. Richard M. Durbin, *et al.* A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 28 Oct 2010.