RN/11/14
14 June 2011

Research
Note

More Mouldy Data: Virtual Infection of the Human Genome

W. B. Langdon and M. J. Arno

Electronic Mail: W,Langdon@cs.ucl.ac.uk
URL: http://www.cs.ucl.ac.uk/staff/W.Langdon/

Abstract

The human genome sequence database contains DNA sequences very like those of mycoplasma molds. It appears such moulds infect not only molecular Biology laboratories but were picked up by experimenters from contaminated samples and inserted into GenBank as if they were human. At least one mouldy EST (Expressed Sequence Tag) has transferred from public databases 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.

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

1 Introduction

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 [1; 2] but it appears only technical problems where considered.

We discovered that the definitive publicly accessible database holding the human DNA sequence has been corrupted in a surprising way. It contains the DNA sequence of a mold [3].

More recently we have discovered a second sequence which is probably not human in the human genome. It appears that the time is ripe for a though check of the NCBI GenBank database.

It appears that not only has the Human DNA sequence been "completely sequenced" [1] 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 mold gene has succeeded in not just moving within its own genome [4] nor only jumping horizontally and crossing the species barrier [5] 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.

2 Computational *in silico* Experiment

The anomalous HG-U133 +2 sequence (GenBank AF241217, probeset 1570561_at) we had previously reported [3] was run against the human genome using Blast [6], at the European Bioinformatics Institute EMBL-EMI with their default settings. 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 in the list. As expected the first match is the query sequence itself (EM_HTG:AF241217). Despite [3] 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: [7].) 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.

3 Discussion

It is well known that mycoplasma contamination is rife in molecular biology laboratories [8]. Many labs are routinely periodically sterilised to counter it. Miller *et al.* [8] said mycoplasma contamination has "potentially major consequences for the diagnosis and characterization of diseases using expression array technology." Nonetheless, using RNAnet[1], 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 [3].

One potential fortuitous side effect of the *in silico* spread of mycoplasma contamination is that the Affymetrix HF-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

[1] `http://bioinformatics.essex.ac.uk/users/wlangdon/rnanet/`

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. With the explosive growth of genomic sequence data available via the Internet, including data from the 1000 genome project [9], it seems time to look again at genomic database quality.

Acknowledgments

Matthew Arno manages the Genomics Centre, King's College London, matthew.arno@kcl.ac.uk.

EPSRC grant EP/I033688/1.

References

[1] Adam Felsenfeld, Jane Peterson, Jeffery Schloss, and Mark Guyer. Assessing the quality of the DNA sequence from the human genome project. *Genome Research*, 9:1–4, 1999.

[2] Rolf Apweiler, Paul Kersey, Viv Junker, and Amos Bairoch. Technical comment to "Database verification studies of SWISS-PROT and GenBank" by Karp et al. *Bioinformatics*, 17(6):533–534, 2001.

[3] Estibaliz Aldecoa-Otalora, William B. Langdon, Phil Cunningham, and Matthew J. Arno. Unexpected presence of mycoplasma probes on human microarrays. *BioTechniques*, 47(6):1013–1016, December 2009. Letter to the editor.

[4] Barbara McClintock. A cytological and genetical study of triploid maize. *Genetics*, 14(2):180–222, 1929.

[5] Tomoichiro Akiba, Kotaro Koyama, Yoshito Ishiki, Sadao Kimura, and Toshio Fukushima. On the mechanism of the development of multiple-drug-resistant clones of shigella. *Japanese Journal of Microbiology*, 4:219–227, Apr 1960.

[6] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped BLAST and PSI-BLAST a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, 1997.

[7] 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.

[8] Crispin J. Miller, Heba S. Kassem, Stuart D. Pepper, Yvonne Hey, Timothy H. Ward, and Geoffrey P. Margison. Mycoplasma infection significantly alters microarray gene expression profiles. *BioTechniques*, 35(4):812–814, October 2003.

[9] A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 28 Oct 2010.