Editorial

Prospects for the automated extraction of mutation data from the scientific literature

Date received (in revised form): 22 September 2010

Recently, Kuipers *et al.*[1] reported an exciting development: the first disease-focused, locus-specific mutation database (LSDB), developed through the automated extraction of mutation data from full-text articles taken from the scientific literature. Although this report clearly constituted a landmark in its field, neither the idea nor the methodology is entirely new, since Horn *et al.*[2] successfully extracted point mutation data reported (mainly in an evolutionary context) in G protein-coupled receptors and nuclear hormone receptors from full-text literature some years ago. Since then a range of different approaches has been attempted, with varying degrees of success.[3–6] Kuipers et al.[1] employed their own search tool, *Mutator*, however, specifically to extract disease-related missense and nonsense mutation data from the scientific literature, with a view to constructing an LSDB for Fabry disease (FMDB) containing mutations in the α-galactosidase (*GLA*) gene. Briefly, these authors employed a disease-oriented PubMed keyword search to identify relevant articles from the literature to be downloaded. The full text of relevant publications was then automatically screened for mutation data, and those mutations found to occur in amino acid residues that appeared to match the GLA protein sequence were selected for inclusion in the database. Although the results appear at first glance to be impressive, they warrant closer inspection.

Using *Mutator*, Kuipers *et al.*[1] identified 367 'unique *GLA* gene mutations' (listed in their Supplementary Table), 108 of which they claimed to be absent from the Human Gene Mutation Database (HGMD[7]). Despite the fact that the authors failed to evaluate *Mutator* with respect to 'standard' performance measures (eg precision and recall), their tool appears to be a significant improvement over previously published methods, particularly those that simply screened PubMed abstracts rather than full text.[6,8–12]. In their comparison with HGMD data, however, Kuipers *et al.*[1] appear to have used the somewhat outdated (but nevertheless freely available) online version of the database (http://www.hgmd.org) rather than the up-to-date subscription version, HGMD Professional (http://www.biobase-international.com/pages/index.php?id=hgmddatabase). We examined the 108 *GLA* mutations claimed to be absent from HGMD, and found that 48 were actually present in HGMD Professional. Of the remainder, seven were listed in HGMD under an alternative mutation type (eg small indels), four still remained unresolved (in that the precise nature of the nucleotide change was unclear or ambiguous in the original report), while 24 were non-Fabry disease false positives (see below). The remaining 25 mutations (~7 per cent of the total number of *bona fide GLA* mutations), reported in 12 different papers, appear to have been inadvertently omitted by HGMD. This is most likely because (i) they were mentioned only briefly within the text of the article concerned and (ii) no hint of the articles'

mutation content could have been gleaned from inspection of article titles, abstracts or keywords. False negatives appear to be less of a problem than false positives, with *Mutator* failing to recognise only nine of the *GLA* mutations logged by HGMD Professional.

The success of *Mutator* in identifying these hitherto latent lesions certainly testifies to the future potential utility of automatic tools designed to search for and extract mutation data directly from full-text articles. Indeed, we are currently exploring ways of incorporating automated full-text searching into HGMD's mutation identification strategy. In terms of the currently available version of the *Mutator* program, however, there would appear to be a significant problem with false positives. As mentioned above, when we carefully examined the 108 identified *GLA* mutations that were claimed to be absent from HGMD, 24 entries (22 per cent) proved to be non-*GLA*/non-Fabry disease false positives. Hence, even the insertion of a step into the search program which was designed to check that all identified mutations matched the GLA protein sequence had not prevented *Mutator* from identifying spurious mutation data (eg mutations in other genes located within article titles cited in the reference list that coincidentally matched the GLA protein sequence etc.). Taken together, it would appear that considerable work remains to be done in order to optimise the performance of the program, even for the missense/nonsense mutation category that constitutes the specific target mutation type for *Mutator*. As it stands, a significant amount of manual validation and curation would still need to be performed in order for the *GLA* mutation dataset to be reliable in terms of its content.

Assuming that these initial difficulties can be overcome, what is the potential of this approach in terms of scaling up the extraction of mutation data for the >3,800 human genes[13] currently known to harbour mutations causing and/or associated with human inherited disease? On the basis of the report of Kuipers *et al.*,[1] we would say that the long-term prospects are likely to be very good, assuming that the search criteria can be suitably refined, using Boolean parameters, so as to exclude the false positives. As the authors discovered, however, each gene/protein is likely to have its own particular false positives to contend with. Thus, the search for *GLA* mutations also pulled out mutations in the Gla (γ-carboxyglutamic acid) domains of various coagulation factors. It follows that significant effort must be devoted to avoiding such false positives on a gene-wise basis. In this context, it is pertinent to point out that there are likely to be a number of categories of mutation that will be very difficult to identify correctly (or alternatively to weed out) in an entirely automated fashion. These are likely to include:

- Somatic mutations, which, unlike germline mutations, are not heritable and hence are not causative of inherited disease. Some types of gene (eg tumour suppressors) are characterised by both germline and somatic mutations and both types of lesion may often be reported in the same paper. Careful reading of the manuscript is required in order to identify the germline mutations unequivocally.
- Mutations introduced by *in vitro* mutagenesis and/or molecular modelling and which have not actually been reported in nature ('experimentally generated mutations') will be difficult to distinguish automatically from genuine disease-associated lesions.
- Mutations that have occurred in an evolutionary context (ie in orthologous proteins in non-human organisms over evolutionary time) will be difficult to distinguish in an automated fashion from human disease-associated lesions.
- Mutations at residues that are abundant within specific proteins (eg Gly residues in the collagens) or which occur at identically numbered amino acid residues in many proteins (eg the initiator methionine codon) are likely to represent significant sources of error (especially of the false-positive kind).
- Mutations other than missense and nonsense mutations will require somewhat different search procedures. Thus, identifying other types of micro-lesion (ie micro-deletions,

micro-insertions, indels, splicing-relevant and regulatory mutations) and different types of gross gene rearrangement (which together constitute >40 per cent of reported mutations causing human inherited disease; see HGMD[7]) will not only have to take account of the DNA sequence of the gene in question, but will also require the adoption of new text-mining techniques.

- Polymorphic missense variants that are neutral with respect to function/clinical phenotype and synonymous (silent) mutations that are of direct pathological significance (eg via an influence on splicing) represent two categories of mutation that will be difficult to respectively exclude from, and include in, a given dataset solely by automated methods.
- Mutations in genes whose proteins are (or have been) subject to different amino acid numbering systems will be difficult to identify unequivocally. Nowadays, protein numbering has been largely standardised, so that the initiator methionine is invariably attributed +1. In the literature, however, numbering systems for one and the same protein frequently differ, depending upon, for example, whether the initiator methionine is given as +1 or −1 or whether the amino acid numbering starts before or after the pre-pro-peptide. Since newly discovered exons can also alter amino acid numbering, numbering schemes tend to change over time and frequently display inconsistencies between different literature reports. Such difficulties are not likely to be insuperable, however, particularly if an up-to-date amino acid reference sequence is used as a standard and the DNA sequence context of the mutation can be captured and used in the validation process.
- Mutations reported only at the amino acid sequence level, and which cannot be unequivocally assigned a single valid nucleotide sequence level alteration, will require further manual curation.
- Mutations that appear at first sight to be genuine but upon closer inspection prove to have been mis-typed by the original authors of the article (a very common example is provided by Glu/Gln transpositions), will in all likelihood represent a continual problem for purely automated search procedures.

For all the above-mentioned reasons, it is, at present, hard to see how the automated collation of mutation data can be wholly accurate and reliable without the subsequent deployment of labour-intensive manual validation and curation steps. Even as it stands, however, it is evident that the automated data extraction tool described by Kuipers *et al.*[1] is likely to represent a very valuable adjunct to the semi-automated data-collation procedures currently employed by both the LSDBs and HGMD. The authors are therefore to be congratulated for the remarkable degree of success that their *Mutator* tool has already achieved at the pilot stage of its development. While the prospect of next-generation tools for the extraction and validation of mutation data is awaited with keen interest, for the time being we concur with the view expressed by Winnenburg *et al.*[14] and Caporaso *et al.*[15] that the most cost-effective and reliable approach to mutation data collection and annotation is currently for automated text-mining methods to be integrated into the manual annotation process, and for manual and automated approaches to be used in concert to mine the biomedical literature for pathological gene lesions.

Peter D. Stenson and David N. Cooper
Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK
Tel: +44 2920 744062; Fax: +44 2920 746551;
E-mail: cooperDN@cardiff.ac.uk

References

1. Kuipers, R., van den Bergh, T., Joosten, H.-J., Lekanne dit Deprez, R.H. *et al.* (2010), 'Novel tools for extraction and validation of disease-related mutations applied to Fabry disease', *Hum. Mutat.* Vol. 31, pp. 1026–1032.
2. Horn, F., Laum, A.L. and Cohen, F.E. (2004), 'Automated extraction of mutation data from the literature: Application of MuteXt to G-protein-coupled receptors and nuclear hormone receptors', *Bioinformatics* Vol. 20, pp. 557–568.

3. Rebholz-Schuhmann, D., Marcel, S., Albert, S., Tolle, R. *et al.* (2004), 'Automatic extraction of mutations from Medline and cross-validation with OMIM', *Nucleic Acids Res.* Vol. 32, pp. 135–142.
4. Caporaso, J.G., Baumgartner, W.A., Jr., Randolph, D.A., Cohen, K.B. *et al.* (2007), 'MutationFinder: A high-performance system for extracting point mutation mentions from text', *Bioinformatics* Vol. 23, pp. 1862–1865.
5. Lee, L.C., Horn, F. and Cohen, F.E. (2007), 'Automatic extraction of protein point mutations using a graph bigram association', *PLoS Comput. Biol.* Vol. 3, p. e16.
6. Krallinger, M., Izarzugaza, J.M., Rodriguez-Penagos, C. and Valencia, A. (2009), 'Extraction of human kinase mutations from literature, databases and genotyping studies', *BMC Bioinformatics* Vol. 10 (Suppl. 8), p. S1.
7. Stenson, P.D., Mort, M., Ball, E.V., Howells, K. *et al.* (2009), 'The Human Gene Mutation Database: 2008 update', *Genome Med.* Vol. 1, p. 13.
8. Yip, Y.L., Lachenal, N., Pillet, V. and Veuthey, A.L. (2007), 'Retrieving mutation-specific information for human proteins in UniProt/Swiss-Prot Knowledgebase', *J. Bioinform. Comput. Biol.* Vol. 5, pp. 1215–1231.
9. Witte, R. and Baker, C.J. (2007), 'Towards a systematic evaluation of protein mutation extraction systems', *J. Bioinform. Comput. Biol.* Vol. 5, pp. 1339–1359.
10. Erdogmus, M. and Sezerman, O.U. (2007), 'Application of automatic mutation-gene pair extraction to diseases', *J. Bioinform. Comput. Biol.* Vol. 5, pp. 1261–1275.
11. Cheng, D., Knox, C., Young, N., Stothard, P. *et al.* (2008), 'PolySearch: A web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites', *Nucleic Acids Res.* Vol. 36 (Web Server issue), pp. W399–W405.
12. Furlong, L.I., Dach, H., Hofmann-Apitius, M. and Sanz, F. (2008), 'OSIRISv1.2: A named entity recognition system for sequence variants of genes in biomedical literature', *BMC Bioinformatics* Vol. 9, p. 84.
13. Cooper, D.N., Chen, J.M., Ball, E.V., Howells, K. *et al.* (2010), 'Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics', *Hum. Mutat.* Vol. 31, pp. 631–655.
14. Winnenburg, R., Wächter, T., Plake, C., Doms, A. *et al.* (2009), 'Facts from text: Can text mining help to scale-up high-quality manual curation of gene products with ontologies?', *Brief. Bioinform.* Vol. 9, pp. 466–478.
15. Caporaso, J.G., Deshpande, N., Fink, J.L., Bourne, P.E. *et al.* (2008), 'Intrinsic evaluation of text mining tools may not predict performance on realistic tasks', *Pac. Symp. Biocomput.* Vol. 13, pp. 640–651.