Prioritising genetic mutations by mining the biomedical literature

Summary

We describe a strategy to use large-scale processing of the biomedical literature to provide context for the interpretation of genetic variants implicated in a disease, as identified through DNA sequencing of patient genomes. Effectively, we are using Big Data, in the form of the over 21 million biomedical research publications, to help understand Small Data, in this case individual genetic mutations. The paradigm we use builds on an approach previously developed for protein functional site prediction [Verspoor *et al*., 2012]. The approach uses text mining in post-processing of predictive results to identify highest-confidence predictions. In brief, this work uses prior knowledge, inferred by text mining of biomedical literature, to prioritize structural variations associated with disease.

Introduction

Interpretation of the effect of a single base change in the human genome is a challenging task. Approaches which assist in this interpretation are becoming increasingly important given that disease studies now involve sequencing of the genome. A typical study will identify thousands of variants, of which only a handful will be relevant to the disease. Furthermore, the small size of medical studies limits the statistical power of any search for significant variants. Therefore, prioritization of variants for further wet lab follow-up is a critical task. Annotation of variants via integration of multiple sources of external evidence regarding their known functional is an important step towards prioritization [Wang et al 2010]. The Oncotator tool[1] aggregates annotations of mutations from various sources such as UniProt's site-specific protein annotations and classes of mutations derived from the Cancer Gene Census[2]. However, it is well known that population of such curated resources is lagging well behind the state of knowledge as represented in the biomedical literature [Baumgartner *et al*., 2007]. To address this, we explore the use of large-scale text mining of that literature to provide a background of evidence of known genetic variant-disease relationships. This background is then used to prioritize variants based on their relevance to the disease study.

Description

As a case study, we apply our method to the annotation of variants discovered via whole-genome sequencing of prostate cancer samples. We focus not only on extracting information from the literature about the role of specific variants in prostate cancer, but also cancer in general, as many underlying disease mechanisms can be shared across cancers. To do this, we restrict our search to the subset of PubMed which is related to cancer, through a simple query for the term "cancer" in the PubMed interface. This literature is processed with a natural language processing toolkit to identify and annotate all mentions of gene names and mutations. Mutations are identified using the EMU tool [Doughty *et al*., 2011], augmented with a search for more generic terms denoting mutation (e.g. "mutation", "deletion", "SNP", "polymorphism") and gene mentions using BANNER [Leaman and Gonzalez, 2008]. These more generic terms are used to identify genes that are

[1] http://www.broadinstitute.org/oncotator/
[2] http://www.sanger.ac.uk/genetics/CGP/Census/

implicated in cancer, even where a specific mutation is not identified. This is a high-sensitivity strategy but likely to be noisy, so such looser associations are kept separate from more specific ones for analysis purposes. The co-occurrence of a gene name and a mutation within an abstract is used to establish a relationship between them. Given our current processing of over 2.7 million cancer-related abstracts identified from PubMed, we have identified well over a million such co-occurrences, involving nearly 250,000 non-normalized gene mentions.

These gene-mutation relationships provide the backdrop for interpreting genetic variations that have been detected from our experimental patient data. We build on the insight from prior work that a mere mention of a specific mutation in a published abstract is evidence of the functional importance of that mutation [Verspoor *et al*., 2012], and use identification of a genetic variant mention in the literature as corroborating evidence for the significance of that variant. This evidence effectively allows us to assign a weight to the variants, and re-rank the variants based on their potential relevance to the disease.

Conclusion

The overarching aim of our work is to identify the genetic variation that appears to be responsible for development of lethal prostate cancer. While we have access to a significant amount of experimental data from individual patients, the interpretation of specific genetic differences between cases and controls is challenging. Our approach gives us a mechanism to exploit the information available in the vast resource of the published biomedical literature to identify genetic variations that have associated prior evidence of functional importance.

References

Baumgartner Jr. WA, Cohen KB, Fox L, Acquaah-Mensah GK, Hunter L (2007). Manual curation is not sufficient for annotation of genomic databases. Bioinformatics, 23:i41-i48.

Doughty, E., Kertesz-Farkas, A., Bodenreider, O., Thompson, G., Adadey, A., Peterson, T., & Kann, M. G. (2011). Toward an automatic method for extracting cancer-and other disease-related point mutations from the biomedical literature. Bioinformatics, 27(3), 408-415.

Leaman, R., & Gonzalez, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. In Pacific Symposium on Biocomputing (Vol. 13, pp. 652-663).

Verspoor KM, Cohn JD, Ravikumar K, Wall ME: Text mining improves prediction of protein functional sites. PLoS One 2012, 7(2).

Wang K, Li M, Hakonarson H (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38(16): e164.