Published in final edited form as:
Stud Health Technol Inform. 2015 ; 216: 1037.

Text Mining and Data Modeling of Karyotypes to aid in Drug Repurposing Efforts

Zachary B. Abrams[a], **Andrea L. Peabody**[a], **Nyla A. Heerema**[b], and **Philip R. O. Payne**[a]

[a]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, Ohio, USA

[b]Department of Pathology, College of Medicine, The Ohio State University, Columbus, Ohio, USA

Abstract

Karyotyping, or visually examining and recording chromosomal abnormalities, is commonly used to diagnose and treat disease. Karyotypes are written in the International System for Human Cytogenetic Nomenclature (ISCN), a computationally non-readable language that precludes full analysis of these genomic data. In response, we developed a cytogenetic platform that transfers the ISCN karyotypes to a machine-readable model available for computational analysis. Here we use cytogenetic data from the National Cancer Institute (NCI)-curated Mitelman database1 to create a structured karyotype language. Then, drug-gene-disease triplets are generated via a computational pipeline connecting public drug-gene interaction data sources to identify potential drug repurposing opportunities.

Keywords

Cytogenetics; Karyotype; Text mining; Drug repurposing

Introduction

Cytogenetic data in the form of karyotypes are commonly used in the diagnosis and treatment of many forms of cancer. Karyotype data are expressed in a text-based form that is not machine-readable. This limits the utility of these data for secondary use and research purposes. Utilizing the International System for Human Cytogenetic Nomenclature (ISCN), we developed a parsing and mapping system that allows karyotype data to be represented and analyzed in a computationally tractable manner. A Loss-Gain-Fusion model (LGF) was created that allowed us to represent each karyotype as a binary vector. Each cytogenetic region is represented three times (loss, gain, and fusion) in the model. We utilized the publicly available Mitelman database as a test-bed for analyses, focusing on problems related to drug repurposing.

Address for correspondence, 350 Lincoln Tower 1800 Cannon Dr. The Ohio State University Columbus, OH 43210 United States

Materials and Methods

Utilizing our computational model and the Mitelman database, we were able to successfully parse 98% of its karyotypes; of those parsed, 89.4% could be mapped into our binary Loss-Gain-Fusion model. We then classified karyotypes based on their disease labels and filtered out all diseases with less than 50 patients. We then selected genetic aberrations present in 20% or more of the population in which the cytogenetic event led to increased gene expression. Subsequently, we identified all genes in the affected region and found drugs that inhibited the function of the overexpressed gene using publicly available drug data in The Drug Gene Interaction Database (DGIdb). We performed a literature search on these results in PubMed selecting diseases and drugs that did not co-occur and where the disease and the gene had co-occurred in at least one PubMed abstract.

Results

We discovered 68,543 triplets containing (1) a disease, (2) an overexpressed gene, and (3) a drug that suppressed that specific gene. From this list, we discovered a total of 69 cancer disease-drug pairs that were not cited as co-occurring in the literature. Given this filtering process where the drug and gene are related, the drug suppressed the gene and the gene was implicated in the disease; it logically follows that the drug should be helpful in treating the disease.

Discussion

Our computational approach serves as a basis for new directions in drug repurposing, leveraging existing and commonly available bio-molecular phenotypic data. In order to validate our results, future laboratory-based testing will be conducted on a sub-set of our findings. The ability to link publicly available data sources is a central component of this work and emphasizes the importance of utilizing such data in conjunction with clinically-generated data sets so as to support in-silico hypothesis generation.

Conclusion

Utilizing publicly available data sets, we generated a list of 68,543 drug-gene-disease triplets, each triplet containing a gene that is up-regulated, a drug that works to suppress or inhibit that up-regulated function, and a disease where the up-regulated gene is an implicated disease agent. This information may play a significant role towards drug repurposing efforts in the 69 diseases. We anticipate that this information will prove useful to researchers in the domain of pharmaceutical drug repurposing and in the treatment of these conditions.

References

1. Mitelman, F.; Johansson, B.; Mertens, F. Mitelman Database of chromosome aberrations and gene fusions in cancer [Internet]. 2014. Available from: http://cgap.nci.nih.gov/Chromosomes/Mitelman/