A New Tool for Prioritizing Candidate Genes

Ayyoub Salmi[*], Romain Philippe[**]
Veronique Blanquet[**] Ahmed Moussa[*]

[*]Systems and Data Engineering Team
University Abdelmalek Essaadi, Tangier, Morocco
[**]Animal Molecular Genetics Unit
University of Limoges, Limoges, France

Abstract. The identification of genes implicated in a disease or in a complex trait from high-throughput experimental studies remains a challenge. Gene prioritization methods play an important role in identifying the most prominent genes in a study. In this paper we present a new tool for prioritizing candidate genes by applying text mining to data related to genes. We tested our tool using genes near a Quantitative Trait Locus (QTL) related to bone weight in cattle, and differentially expressed genes related to Parkinson Disease.

1 Introduction

The recent advances in sequencing technologies led to an increase in organisms with a completely sequenced and, more importantly, annotated genome Reuter et al. (2015). However, the task of identifying genes that are involved in a particular disease or related to a trait remains challenging.

Candidate gene studies evaluate genetic variation in the region of genes that are physiologically suggested (candidates) to be involved in disease pathogenesis Holloway and Prescott (2017) or related to a desired trait.

In the present article, we describe a new ranking tool that uses text mining to prioritize candidate genes following some criteria using gene expression, gene function, homology, knock-out studies and a survey of literature.

2 Methods

In order to use the text mining approach, we first collated gene related information from different sources. For this purpose we created a database containing 56813 human genes including coding genes, non coding genes and pseudogenes. We used Elasticsearch, a distributed full-text search engine to store and interrogate those documents.

Since we are using a text mining approach, a set of keywords describing the the disease or phenotype under study must be provided. The drawbacks of using keywords can be observed when expressing complex relations or when making novel predictions. To overcome those limitations, we introduced a new strategy to find relevant keywords to those initially provided. Assuming that words contained in a document describing a gene have a certain frequency of appearance, a potential keyword is defined as a word with higher than expected frequency.

Relevant genes are first highlighted using a scoring system described by the following formula:

$$\log(1 + F_{t,d}) \times \log(\frac{2 \times N}{1 + F_t}) \tag{1}$$

Where $F_{t,d}$ is the number of evidences containing the keyword t in gene d, N is total number of genes and F_t is the number of genes where keyword t appears.

Genes with a score different than zero are used to find a new set of keywords relevant to the ones initially used. The acquired sets of words for each gene are then compared to produce a final set of words that represents a profile of keywords. A second round of text mining is then performed.

At the end, the user is presented with a report containing relevant genes and their corresponding scores, and also the different evidences that were found while highlighting the keywords used to find them.

3 Results

We have implemented our approach for detecting candidate genes in java.

For the purpose of testing the application, we run two tests using two different datasets: the first one contains genes from a QTL region and the second one contains differentially expressed genes in patients with Parkinson Disease.

4 Conclusion

In this paper we presented a new tool for Prioritizing candidate genes involved in a disease or related to a desired trait using text mining.

The adopted approach is based on the assumption that a gene is more likely to be a candidate if it has a known physiological role in the phenotype of interest and it affects the trait in question based on studies of knock-outs in other species.

For the purpose of testing our tool, we used two set of genes: the first one comprises of a list of genes located near a QTL region in cattle genome related to bone weight, and the second one contains differentially expressed gene related to Parkinson Disease.

References

Holloway, J. W. and S. L. Prescott (2017). *The Origins of Allergic Disease*, pp. 29–50. Elsevier.

Reuter, J. A., D. V. Spacek, and M. P. Snyder (2015). High-throughput sequencing technologies. *Molecular Cell 58*(4), 586–597.