A quantitative analysis of causal and associative events involving genes and proteins

Phoebe M. Roberts
Computational Sciences Center of Emphasis
Pfizer Inc., Cambridge MA

Biomedical research relies on indirect readouts to test hypotheses and reach conclusions from a set of experimental results. For instance, to unequivocally demonstrate that a gene region from a disease association study contains a disease-causing mutation requires the following: identifying the segregating mutation, demonstrating that it affects protein levels or function, and showing how the altered protein levels or function disrupt biological processes and lead to disease ((Musunuru, Strong et al. 2010)). These high-impact causal events are the culmination of sets of experiments that collectively refute alternative hypotheses and uncover the mechanistic underpinnings of disease. In the absence of support for causality, many findings are reported as associative when the data show a correlation between a gene product and disease (e.g. a biomarker for disease) or a genetic association.

To better understand how causal and associative gene and protein events are represented in biomedical abstracts, a set of causal and associative verbs was assembled by using known representative relationships to retrieve verbal connectors (referred to in (Rodriguez-Esteban, Roberts et al. 2009). For example, vascular endothelial growth factor (VEGF) is a well-known angiogenesis-inducing protein, and inhibiting signaling through VEGF has led to several marketed therapies for the treatment of cancer and age-related macular degeneration (Ferrara 2009). A straightforward "VEGF [verbal relationship] angiogenesis" query retrieves verb phrases that reflect causal, inductive relationships, such as "induces", "stimulates", and "is critical for". Causal inhibitory and associative verb phrases were collected using similar well-studied relationships, and they were consistent with other biomedical verb classification studies (Rebholz-Schuhmann, Jimeno-Yepes et al. 2010).

With a gene dictionary and set of relevant verbs in hand, 334 noun phrases that followed the verb phrases were collected and classified into categories that were determined by an initial review of the results: biological processes, state changes of genes and gene products, phenotypes, diseases, and an "other" category for noun phrases that did not fit elsewhere. The four categories together were sufficient to classify 95% of the results. The difference in distribution of term classes between causal vs. associative verbs was striking; biological processes and state changes of genes accounted for 72% of the causal events, vs. 15% of the associative events. In contrast, diseases and phenotypes were found in 22% of the causal relationships, vs. 81% of the associative relationships.

This work has identified distinct patterns in causal vs. associative relationships involving genes and gene products. Prevalent terms from causal gene relationships have the potential to provide intermediary connections in gene-disease relationships.

BIBLIOGRAPHY

Ferrara, N. (2009). "VEGF-A: a critical regulator of blood vessel growth." Eur Cytokine Netw **20**(4): 158-163.

Musunuru, K., A. Strong, et al. (2010). "From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus." Nature **466**(7307): 714-719.

Rebholz-Schuhmann, D., A. Jimeno-Yepes, et al. (2010). "Measuring prediction capacity of individual verbs for the identification of protein interactions." J Biomed Inform **43**(2): 200-207.

Rodriguez-Esteban, R., P. M. Roberts, et al. (2009). "Identifying and classifying biomedical perturbations in text." Nucleic Acids Res **37**(3): 771-777.