Large-scale Automated Reading with Reach Discovers New Cancer Driving Mechanisms

Marco A. Valenzuela-Escárcega[1], Özgün Babur[2], Gus Hahn-Powell[3], Dane Bell[3], Thomas Hicks[1], Enrique Noriega-Atala[4], Xia Wang[5], Mihai Surdeanu[1], Emek Demir[2], Clayton T. Morrison[4]

[1]Dept. of Computer Science, University of Arizona, Tucson, USA
[2]School of Medicine, Oregon Health and Sciences University, Portland, USA
[3]Dept. of Linguistics, University of Arizona, Tucson, USA
[4]School of Information, University of Arizona, Tucson, USA
[5]Dept. of Molecular and Cellular Biology, University of Arizona, Tucson, USA

***Abstract*—PubMed, a repository and search engine for biomedical literature, now indexes more than 1 million articles each year. This exceeds the processing capacity of human domain experts, limiting our ability to truly understand many diseases. We present Reach, a system for automated, large-scale reading of biomedical papers that can extract descriptions of biological processes at a mechanistic level of detail and with relatively high precision. We demonstrate that combining the extracted pathway fragments with existing biological data analysis algorithms that rely on curated models helps identify and explain a large number of previously unidentified mutually exclusive altered signaling pathways in seven different cancer types. This work demonstrates that combining curated "big mechanisms" with extracted "big data" can lead to a causal, predictive understanding of cellular processes and unlock important downstream applications.**

Keywords: machine reading, biological data analysis, hybrid human-machine models

In the period of 2004–2013, over 7.3 million journal articles were added to PubMed [7], and the rate is now over 1 million articles per year. At the same time, a typical large-scale profiling effort now produces petabytes of data – and is expected to reach exabytes within the near future [3]. Unfortunately, most of the mechanistic knowledge in the literature is not in a computable form and therefore remains largely hidden. Existing biocuration efforts are extremely valuable for solving this problem, but they are outpaced by the explosive growth of the literature. For example, we estimate that public pathway databases such as Pathway Commons[1] capture only 1–3% of the literature, and the gap widens every day.[2]

This gap severely limits the value of big data in biology. As a concrete example, consider the detection of "driver" mutations in cancer. One widely recognized observation is that, given a cohort of patients, some driver alterations will exhibit a mutually exclusive pattern. That is, the number patients that have both drivers will be smaller than what is expected by chance. This often happens because these alterations unlock the *same* cancer driving pathways and the positive selection of one diminishes substantially when the other is present. In other words, "one is enough." Prior pathway knowledge can be used to improve the accuracy of these methods by limiting the search space and reducing the loss of statistical power due to multiple hypothesis testing correction. It also provides mechanistic explanations of the observed correlations [1]. Recall, however, can be low due to the aforementioned database coverage issues. Researchers are thus faced with a choice between no-prior, high coverage methods that do not provide mechanistic explanations or low-coverage, prior-based methods that may overlook some key events.

We propose a natural language processing (NLP) approach that captures a system-scale, mechanistic understanding of cellular processes through automated, large-scale reading of scientific literature, and demonstrate that this approach leads to the discovery of novel biological hypotheses for multiple cancers. We call our approach Reach (REading and Assembling Contextual and Holistic mech-

[1]`www.pathwaycommons.org`
[2]Internal analysis of the Pathway Commons team.

anisms from text).

Reach is a hybrid statistical and rule-based approach, with its core consisting of compact grammars for the recognition of cellular processes. These grammars were developed using the Odin information extraction framework [4, 5, 6]. In all, we recognize 16 event types that follow the BioPAX representation [2], with a relatively small grammar of approximately 150 rules. This focus on grammar compactness is important for two reasons. First, it guarantees that the overall model is *interpretable*, i.e., it can be easily understood, modified, and extended by domain experts. And second, this compact grammar can be applied efficiently, permitting high-throughput processing. In an independently administered evaluation[3], Reach was found to approach human precision at a throughput capable of reading the entire open source biomedical literature within days.

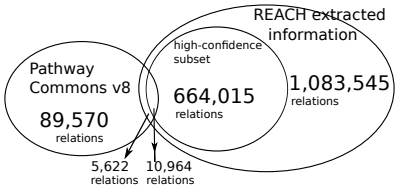

Fig. 1. Reach data is about 17 times larger than the size of Pathway Commons, and they have little overlap. A high confidence set in Reach i.e., relations with more than one literature reference, is about two thirds of this data.

The contributions of this work are two fold. First, we demonstrate that Reach-extracted pathway fragments improve the inference capacity of existing biological data analysis algorithms that already benefit from large curated models ("big mechanisms"). Specifically, we extended the Pathway Commons human-curated pathways with fragments extracted by Reach from all papers in the Open Access subset of PubMed (1,046,662 papers as of June 2015) (Fig. 1). Using this combined prior network we were able to identify previously unidentified, but highly statistically significant mutually exclusively altered signaling modules in TCGA cancer datasets using the Mutex algorithm [1] (Fig. 2).

Second, a manual evaluation of these modules by an external cancer researcher reveals that, despite the inherent noise in machine reading, 65% of the hypotheses proposed by Mutex+Reach are indeed correct according to the literature. Further, a simple redundancy filter that keeps Reach extractions only if they are seen at least twice in the literature increased this accuracy to 80%. This demonstrates that our approach systematically and incrementally increases coverage of prior, curated networks using NLP strategies, and, we believe, is valuable for molecular tumor boards and other cases where one needs to combine system-scale data with the knowledge in the literature.

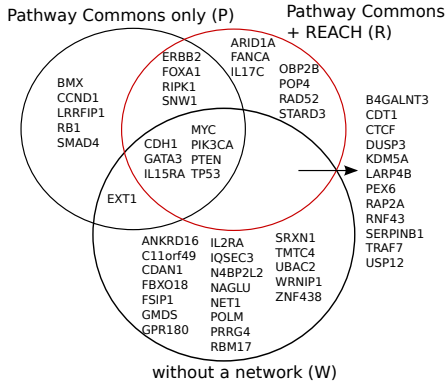

Fig. 2. Using Reach extracted information allows Mutex to detect 7 new "driver" genes for breast cancer which are not detected using Pathway Commons only or without using any network. We observed similar results for 7 cancers in the TCGA dataset.

References

[1] Ö Babur, M Gönen, B A Aksoy, N Schultz, G Ciriello, C Sander, and E Demir. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome biology*, 16(1):45, 2015.

[2] E Demir, M P Cary, S Paley, K Fukuda, C Lemer, I Vastrik, G Wu, P D'Eustachio, C Schaefer, J Luciano, et al. The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, 28(9):935–942, 2010.

[3] Z D Stephens, S Y Lee, F Faghri, R H Campbell, C Zhai, M J Efron, R Iyer, M C Schatz, S Sinha, and G E Robinson. Big data: astronomical or genomical? *PLoS Biol*, 13(7):e1002195, 2015.

[4] M A Valenzuela-Escárcega, G Hahn-Powell, T Hicks, and M Surdeanu. A domain-independent rule-based framework for event extraction. In *Proceedings of the 53rd Annual Meeting of the ACL: Software Demonstrations*, pages 127–132, 2015.

[5] M A Valenzuela-Escárcega, G Hahn-Powell, and M Surdeanu. Description of the Odin event extraction framework and rule language. *CoRR*, abs/1509.07513, 2015.

[6] M A Valenzuela-Escárcega, G Hahn-Powell, and M Surdeanu. Odin's runes: A rule language for information extraction. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 322–329, may 2016.

[7] K Z Vardakas, G Tsopanakis, A Poulopoulou, and M E Falagas. An analysis of factors contributing to pubmed's growth. *Journal of Informetrics*, 9(3):592–617, 2015.

[3]Conducted in the DARPA Big Mechanism program (`www.darpa.mil/program/big-mechanism`).