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 mechanistic descriptions of biological processes 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 shows that combining curated "big mechanisms" with extracted "big data" can lead to a causal, predictive understanding of cellular processes and unlock 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 (1), and the rate is now over 1 million articles per year. Unfortunately, most of the mechanistic knowledge in the literature is not in a computable form and therefore remains 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 Commonscapture only 1–3% of the literature, and the gap widens every day.[1]

This gap severely limits the value of big data in biology. For example, some "driver" mutations in cancer exhibit a mutually exclusive pattern in a given cohort of patients. That is, the number of patients with 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. The Mutex algorithm (2) searches for groups of genes such that the alterations are mutually exclusive, and each gene in the group significantly contributes to this pattern. Pathway knowledge improves Mutex's accuracy 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. 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 without 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 mechanisms 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 recog-

[1]Internal analysis of the Pathway Commons team.

[2]Inspired by NLP literature, we use "event" to indicate an interaction between multiple participants.

nize biological entities (e.g., genes, proteins, protein families, simple chemicals), events[2] that operate over these biochemical entities (e.g., biochemical reactions), and nested events that operate over other events (e.g., catalysis). These grammars were developed using the Odin information extraction framework (3–5). In all, we recognize 16 event types that follow the BioPAX representation (6), 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, i.e. biologists. And second, this compact grammar can be applied efficiently, permitting high-throughput processing.

The Reach architecture is implemented as a cascade of automata that recognize increasingly complex biomedical phenomena, as illustrated in Fig. 1.

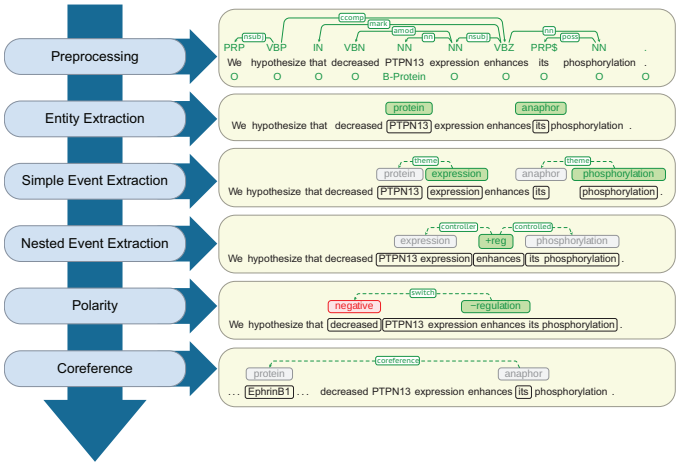

Fig. 1. A walkthrough example of the cascaded Reach architecture.

The Reach pipeline starts with preprocessing, splitting the text into sentences and further into words. Then, the text is annotated with parts of speech (e.g., `NN` indicates common noun), syntactic dependencies (e.g., `nsubj` captures the relation between a verb and its nominal subject), and named entity labels (e.g., `B-PROTEIN` indicates that the corresponding token is the beginning of a protein's name).

Next, Reach searches the preprocessed text. First, it extracts entity mentions (e.g., *PTPN13* as a `Protein` mention), including anaphoric mentions (e.g., *its*) to be resolved later. Second, Reach searches for events that operate directly on these entities, which we call simple events. These rules apply either over word (and part-of-speech) sequences or the dependency graph. For example, the *phosphorylation* event is extracted in this step. Third, Reach recognizes nested events, i.e., those with other events as arguments, such as the positive regulation in the example sentence. All three steps are implemented using an Odin grammar.

After the rule-based extraction, Reach applies further deterministic steps. It first detects event negation using the dependency graph, combining the negative *decreased* and positive *enhances* to produce a negative regulation. Then, it detects that the regulation is hypothesized using the dependency between *hypothesize* and the regulation. Finally, a deterministic coreference-resolution system (7) determines that *its* corefers with a previously mentioned entity *EphrinB1* (rather than *PTPN13*, for example).

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. Participating systems extracted mechanistic information from a thousand papers about the Ras signaling pathway over the course of a week. Three metrics were used to evaluate the participating systems: *throughput*, the estimated number of interactions produced per day; *generous precision*, the proportion of interactions that were considered useful by the expert panel; and *strict precision*, the proportion of interactions that were completely correct. Four consortia, each one potentially containing multiple teams, participated in the evaluation. Team 4 was a consortium between Reach and another group (4(B)). The results are summarized in Table I.

TABLE I. BIG MECHANISM EVALUATION

Team	Throughput	Strict Precision	Generous Precision
Team 1	110	54.69%	62.50%
Team 2	975	35.24%	42.07%
Team 3	242	56.76%	75.68%
Team 4	944	50.36%	66.42%
Reach	760	58.76%	74.23%
Team 4(B)	189	30.00%	47.50%

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[4] 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. 2).

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

[4]`http://www.pathwaycommons.org/`

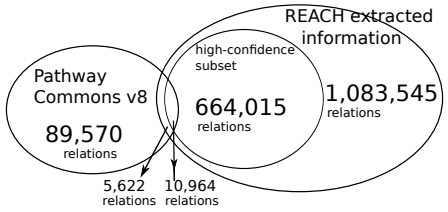

Fig. 2. The Reach output is about 12 times larger than the size of Pathway Commons. We conjecture that the small overlap is caused by the fact that the Reach interactions are extracted from open-access publications, whereas Pathway Commons pathways come mostly from other, paywalled publications. The high-confidence subset is of relations that were found in more than one paper.

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 described above. Fig. 3 shows Mutex groups for TCGA breast cancer, and Table II summarizes the findings for all enhanced cancer studies in TCGA. R represents the Mutex configuration using the combined Reach + Pathway Commons network, P the Mutex configuration using only the Pathway Commons network, and W the Mutex configuration uninformed by any supporting network.

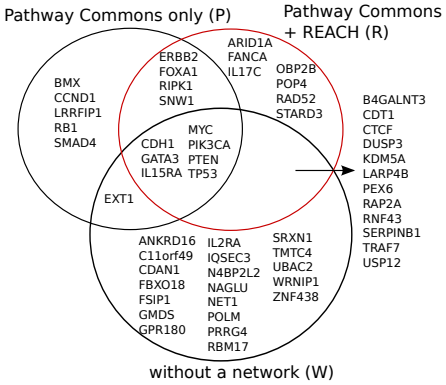

Fig. 3. Using Reach-extracted information allows Mutex to detect 7 new "driver" genes for breast cancer which are not detected using Pathway Commons only or without using any network. We observed similar results for multiple cancers in the TCGA dataset.

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.

TABLE II. MUTEX+REACH ANALYSIS OF TCGA

Cancer study	*R*	P	W	*R* − P − W	*R*W − P
BLCA	2	2	6	0	0
BRCA	30	17	40	7	12
CESC	5	6	7	0	0
DLBC	0	5	0	0	0
GBM	23	14	40	3	7
HNSC	26	23	25	3	2
KICH	0	0	6	0	0
LAML	2	2	2	0	0
LGG	26	12	51	0	14
LIHC	12	17	16	0	0
LUAD	14	16	11	1	0
OV	7	11	7	2	0
PAAD	22	7	17	10	5
SARC	15	22	25	0	0
THCA	9	11	12	0	0
UVM	2	3	34	0	0

REFERENCES

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

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

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

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

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

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

7. Bell, D., Hahn-Powell, G., Valenzuela-Escrcega, M.A. and Surdeanu, M. (2016) Sieve-based coreference resolution in the biomedical domain. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. pp. 177–183.