Data and text mining

NLPCORE – Computer Aided Discovery of Annotations of Bioentities & their Relationships in Articles

Varun Mittal[1], Alexander V. Ratushny[2,3], John D. Aitchison[2,3] and Naveen Garg[1,*]

[1]NLPCORE, Seattle, WA, USA, [2]Center for Infectious Disease Research (formerly Seattle Biomedical Research Institute), Seattle, WA, USA, [3]Institute for Systems Biology, Seattle, WA, USA

[*]To whom correspondence should be addressed.

Abstract

Summary: NLPCORE is a life sciences researchers' workbench built upon a deep neural network search technology that provides contextually relevant biomaterials, their interactions and references to advance fundamental research, drug discovery, or personalized healthcare recommendations. Using NLPCORE's intuitive color-coded visualizations, filters, and self-pruning graphs, researchers identify relevant bioentities and their relationships in published literature to establish well-informed hypotheses for their experiments. We evaluated NLPCORE performance to find known protein-protein and experimentally discovered host gene-virus interactions in a high throughput manner using nearly 1.25 million PubMed Central Open Access articles. NLPCORE discovered selected interactions with >99% recall rate searching 5 levels deep in its network graph. We present our methods and validation results in greater detail in Supplementary Materials.

Availability: http://www.nlpcore.com

Contact: naveengarg@nlpcore.com

Supplementary information: Supplementary details on our methods and data used are available at http://www.nlpcore.io/?attachment_id=110

1 Introduction

With enormity of existing and ever-increasing corpus of published research, it is virtually impossible for life sciences researchers to manually review, extract and rate significance of relevant articles to select and contextualize a handful of candidate proteins, genes, viruses or other bioentities for their experiments. Moreover, currently available tools focus on reporting protein-protein interactions, post-translational modifications, gene expression data, or expert-curated functional information (Franceschini *et al.*, 2013) and fall short on discovering many other interactions e.g. viral-host interactions. Researchers also cannot easily filter results using both quantitative information (such as biological properties) and qualitative information extracted from specific annotations in manuscripts (such as linkages across specific proteins, genes and reagents). Given these challenges, this area has attracted many text mining approaches from trained classifiers, natural language processing, machine learning to relationship generation (Cohen *et al.*, 2005).

NLPCORE (Fig. 1A) is a broad search platform based on deep neural network techniques that extracts bioentities (organisms, cell types, genes, RNAs, proteins, metabolites and others) annotations in scientific publications along with associated and relevant information such as methods and materials, specific researchers, related biological entities and reagents as well as results of experiments conducted with their relative significance (based on citations, results, proximity and frequency of references). NLPCORE enables researchers to discover new interactions, more precisely rank, filter and select candidate bioentities from millions of prior research and clinical trial documents. We envision it to become an integral part of biomedical research and experimental design for cell and molecular biology and beyond. Using available research from PubMed, Bioentity dictionaries, references and interaction databases NLPCORE collects references of attributes and relationships for bioentities to ensure highly accurate and precise recognition of their annotations along with their cross-references in the entire corpus. With integration of manual curation, collaboration tools, community ratings over time NLPCORE can potentially become an indispensable workbench for researchers to formulate more objective and informed hypotheses, to select materials and reagents to experiment with and to find potential collaborators.

2 Methods

Using NLPCORE's user-interface researchers visually explore color-coded relationship graphs be it protein-protein, protein-gene, gene-gene, viral-host, or other types of associations, along with their annotation references (Fig. 1A). They apply filters on entity names, search terms and extracted conceptual phrases (graph nodes) in a consistent and intuitive fashion to continually explore and discover more relationships. They can also dial how deep relationships (graph edges) are explored and what criteria to apply to rate, rank and filter them including algorithm's confidence score, article's publication date, or journal's impact factor. NLPCORE allows researchers to therefore more precisely identify and situate candidate biomaterials into their experiments' hypothesis.

Despite great advances in text mining techniques, classification aids (Kim *et al.*, 2003) and frameworks (Settles, 2005), there remains a large gap primarily attributed to training datasets that require labor intensive manual curation. In addition, many clustering or classification approaches using simple grammar rules falter on research papers written in a more complex scientific vocabulary. Another challenge is the uneven distribution of concepts. A standard text mining approach may expose relations around the known biological terms referenced in an article but might not iterate beyond to a greater depth. This results in most of the discovered relations to be biased towards trained and known datasets. Instead, NLPCORE employs a unique deep neural network technology (Craven and Shavlik, 1997) to identify meaningful relationships consisting linguistically (parts of speech - noun, pronoun, verb, adjective, adverb, etc.) and statistically (term frequencies, proximity scores, distribution of terms across corpus and across one another, etc.) relevant phrases. It is inherently designed to be corpus agnostic and we can traverse it deeply at search time without a significant compute cost. We do so by keeping track of variances (Rovatti and Mazzini, 2008) amongst word relations while doing all the computations. The variance helps us identify cluster of meaningful concepts. The ones that do not change frequently, often appear in the center of the cluster(s) (in contrast with a static trained or known dataset). This approach efficiently tags bioentities and clusters them with known types allowing for application of additional rules and known dictionaries to find even more interactions. The graph is also kept readily traversable with a secondary full-text index on top. This well connected graph in NLPCORE is continually evolving with new tags, relationships, and new scores (linguistic attributes, proximity/term frequency, impact factor or user interactions). NLPCORE can entertain various filters on entity nodes (such as properties from biodictionaries) and various controls on their relationship edges (such as confidence scores, publication date of articles, impact factor of journals, most linked or least linked nodes, etc.). Users therefore can easily explore the entire search space without needing to learn new search syntax, or read through vast corpus, instead using simple filter or slider controls discover new interactions, pathways, transformation or linkages across various bioentities, reagents and researchers.

From the ground-up, NLPCORE's components such as data collection, extraction, computational processing and storage are built cost appropriately and as independently scalable units to run on an on-demand Hadoop cluster. We have successfully built reference implementations on small micro-servers, Amazon, Google, Microsoft, Private cloud as well as Onsite Sandboxed infrastructures processing millions of articles. More recently, we have employed massively parallel computing techniques (by offloading them from CPU to GPU hardware) to search more deeply in our graph and have found excellent results in our pilot studies on diverse protein kinase interactions (Franceschini *et al.*, 2013) (Fig. 1.B).

In summary, NLPCORE is a significant improvement and complements many of the existing biomedical text mining tools. It offers:

- A continually evolving deep network of meaningful concepts and their interrelationships;
- An improved linguistic parser for complex paragraphs coupled with numerous statistical models to extract most meaningful concepts;
- Integration with biodictionaries, interaction databases, existing tools, meta-tags and heuristics to cluster concepts as bioentities, reagents, institutions and researchers;
- Ability for users to search with simple keywords and visually explore color coded entity graph with annotation references at any depth using filters on bioentities, reagents, researchers, search terms, concepts or by applying sliding ranges on relationship scores, article publication dates or journal impact factors.

3 Results

We evaluated NLPCORE performance to find known protein-protein and/or newly experimentally discovered host gene-virus interactions using nearly 1.25 million PubMed Central Open Access articles. Fig. 1B-C summarizes the NLPCORE performance test results. Fig. 1B shows number of interactions for nearly 170 commonly used proteins (Franceschini *et al.*, 2013) discovered from articles. As we traverse deeper in the relationship graph, our true positive and recall rates increase substantially with a smaller increase in false positives (Supplementary Materials). Fig. 1C provides comparison of references of protein-virus interactions in terms of their published paper counts against the experimentally discovered (Schoggins *et al.*, 2014) viral-host interactions or non-interactions. Our results not only corroborate manually curated and experimental data but also provide novel interactions (that are not commonly known or easily discoverable) in the existing publications (Supplementary Materials).

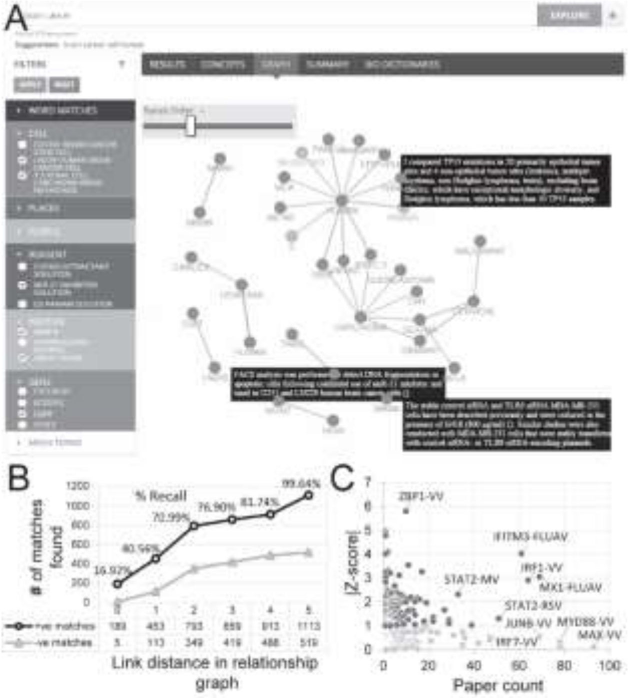

Fig. 1. NLPCORE web-based platform. (A) Examples of color-code filters and relationship graphs. (B) Count of protein-protein interactions discovered up to 5 links deep in relationship graph with increasing recall rates; (C) Comparison of NLPCORE derived (x-axis; paper count) and experimentally discovered (y-axis, absolute values of Z-scores) host gene-virus interactions. Experimental data represent the results of the ectopic expression assay used to screen a library of over 350 human interferon-stimulated genes (ISGs) for effects on 14 DNA, positive- or negative-sense single-stranded RNA viruses. Z-scores represent virus replication values of ISG screen data sets. VV, FLUAV, MV, and RSV are vaccinia, influenza A, measles, and respiratory syncytial viruses, respectively.

Acknowledgements

We are thankful to the guidance and unlimited access to research labs and ongoing projects, provided by Center for Infectious Disease Research in Seattle, WA and University of Washington in Tacoma, WA.

Funding

This work has been supported by the NLPCORE LLC. Ratushny A. and Aitchison J. were supported by grants P50 GM076547 and P41 GM109824 from the National Institutes of Health.

Conflict of Interest: none declared.

References

Cohen A.M., Hersh W.R. (2005) A survey of current work in biomedical text mining. Brief. Bioinform. 6, 57–71.

Craven M.W. and Shavlik J.W. (1997) Using neural networks for data mining. Future generation computer systems. 13, 211–229.

Franceschini A., et al. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 41, D808–815.

Kim J.D. *et al.* (2003) GENIA corpus—a semantically annotated corpus for biotextmining. Bioinformatics. 19, i180–i182.

Rovatti R, Mazzini, G. (2008) On the Nearest Neighbor of the Nearest Neighbor in Multidimensional Continuous and Quantized Space. IEEE Transactions on Information Theory. 54, 4069–4080.

Schoggins J.W., et al. (2014) Pan-viral specificity of IFN-induced genes reveals new roles for cGAS in innate immunity. Nature. 505, 691–695.

Settles B. (2005) ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics. 21, 3191–3192.