RESEARCH

Linked Annotation Networks: Developing an Infrastructure for Inference

Tiffany J Callahan*† and Lawrence E Hunter

*Correspondence:
tiffany.callahan@ucdenver.edu

University of Colorado Denver Anschutz Medical Campus, 12800 E. 19th Avenue, Aurora, CO, USA

Full list of author information is available at the end of the article

†Equal contributor

Abstract

In the field of biology, networks facilitate the reconstruction and visualization of complex biological processes. Within a network, nodes represent biological components (i.e., proteins or genes) and edges denote the relations (i.e., physical interactions or regulatory relationships) between nodes. Inference of the network structure or 'wiring diagram' provides information regarding the functional relationships between network elements [1]. For example, an expression network (nodes: genes/proteins and edges: regulatory relationships) can be used to predict which genes/proteins are most likely to influence each other and by what mechanism (i.e. transcriptional regulation or posttranslational modifications, etc.) [1]. Linked literature annotations are well suited for network inference. We hypothesize that network inference of linked annotations will accelerate current annotation efforts through the prediction of future annotations, will enhance current data analysis efforts, and will help with the characterization and tracking of current literature. To test this hypothesis an infrastructure for developing and evaluating network representations from linked annotations is required. Network representation of linked biomedical annotations has yet to be fully explored and is the goal of the proposed work.

Keywords: Linked Annotation; Biomedical Data; Networks; Network Inference

Background

The volume of health data produced worldwide is expected to reach 25,000 petabytes by 2020, a 50-fold increase from the amount of data generated in 2012 [2]. The generation and analysis of these data has already facilitated an exponential accumulation (>3,000 published articles/day) of peer-reviewed literature [3,4]. To help synthesize this information, the natural language processing and biomedical research communities have developed a collection of manually annotated text corpora. While the breadth of knowledge represented in existing corpora is extensive, the depth of represented entities and their relationships is limited [5]. As highlighted in prior Biomedical Linked Annotation Hackathon (BLAH) proposals, integrating annotated corpora with biomedical ontologies can improve the coverage of represented entities and their described relationships [6], but complete integration of these resources is often impracticable.

Grounded in the field of graph theory, networks help to decipher complex processes [7] and have been utilized extensively in biomedical research to represent every-

thing from cell signaling [8] to drug-drug interactions [9]. The structure of linked annotations is conducive for network representation, but the potential network configurations that can be represented from these data are exponential. Understanding the benefits and limitations of different network representations for these data is an important first step to evaluating their utility for network inference.

As a proof-of-concept, three approaches for generating networks using linked open data (i.e. PubAnnotation and the Gene Ontology (GO)) are proposed. For each approach, the following test sentence from a CRAFT Treebank annotated text (10) was used:

Bone morphogenetic protein (BMP) family members, including BMP2, BMP4, and BMP7, are expressed throughout limb development.

Syntactic Treebank Networks

The first network generation method leverages syntactic Treebanks (Figure 1). This network was constructed by linking the annotated parts of speech from the test sentence according to the order of appearance in the Treebank hierarchy. The nodes in this network represent the parts of speech and two nodes are connected if they occur within the same word or part of speech.

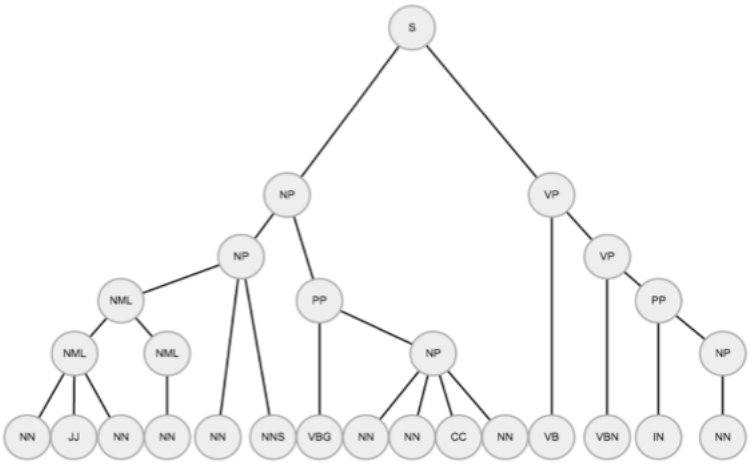

Figure 1 Syntactic Treebank Network.

Semantic Annotation Networks

Semantically annotated text can be used to construct an ontology annotation network (Figure 2). In this network, nodes represent the concept ancestors of the semantically annotated text entities. The ontology entity relations determine the edge between two nodes. Figure 2 portrays the network representation of GO terms for anatomical structure morphogenesis (purple), gene expression (light blue), and limb development (yellow node). With this network representation, the co-occurrence of nodes between each of the test sentence GO terms could be investigated to provide insight into their semantic similarity at different levels of the GO ontology.

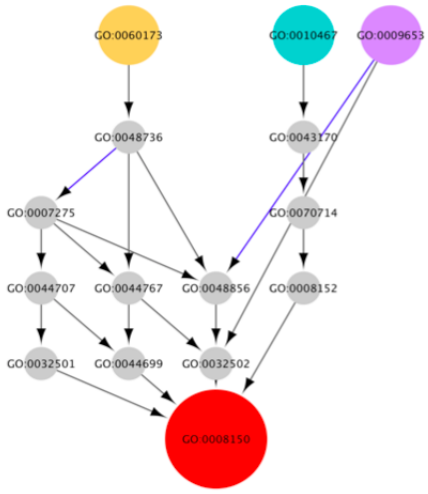

Figure 2 Semantic Annotation Networks. The root node (Biological Process) is colored red. GO ontology terms mapped to text in annotations: limb development (yellow), gene expression (light blue), and anatomical structure morphogenesis (purple). Edge colors represent entity relations from the GO: is a (gray) and part of (blue).

Sequential Concept Networks

Sequential networks are built by aligning sequences of annotated terms and/or concepts (Figure 3 and 4). From syntactic annotations, networks can be generated by linking words, phrases and sentences, or even paragraphs or full document sections. The three sub-networks in Figure 3 were built from the network shown in Figure 1 and represent the sequence of nodes within path length two or fewer from three separate spans of text (NML: Bone morphogenetic protein, PP: including BMP2, BMP4, and BMP7, and PP: throughout limb development).

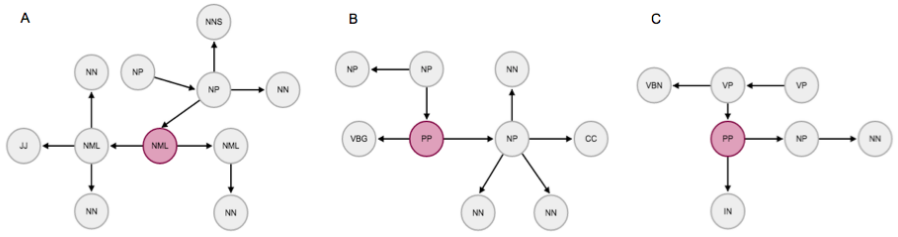

Figure 3 Syntactic Sequential Concept Networks. Figure 3A covers words with the span of 92-139; Figure 3B covers words with spans 141-171; and Figure 3C covers words with the span of 173-214.

Semantic annotations can be sequentially linked by aligning the concepts according to the order of occurrence in text or by leveraging a mapped ontology hierarchy (Figure 4). This network illustrates different ways to model the text surrounding the ontology terms anatomical structure morphogenesis and limb development. For these networks, nodes represent words or phrases and two nodes are connected if they represent a term/concept. Dashed lines demonstrate how the different sequential sub-networks could be connected to form a larger network representation.

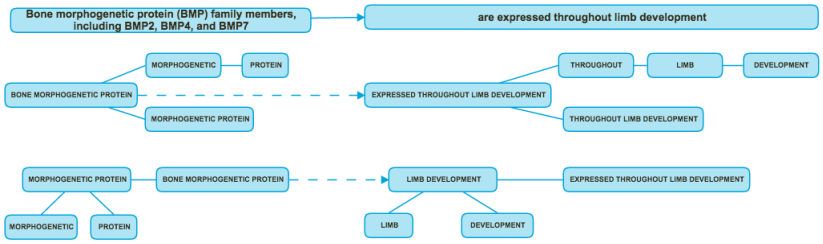

Figure 4 Sequential Concept Networks. Solid lines represent direct connections, dashed lines are used to demonstrate potential adaptions that cane be made to combine the different network representations.

Proposed Work

We propose the following work for the hackathon:

- Develop code that generates network representations for an annotated text from three PubAnnotation corpora (CRAFT, DisGeNET, and LocText).
  - For each text, Treebank, semantic annotation, and sequential concept network representations will be developed using open source software.
  - For each text, generate network representations for multiple levels of annotation (i.e. words or phrases, sentences, paragraphs, and full document sections).
- Evaluate the utility of generated network representations by:
  - Creating visualizations for each network using open source visualization tools (i.e. Cytoscape, Gephi).
  - Providing descriptive network statistics (e.g. number of nodes/edges, average degree, centrality, assortativity, and average path length).
  - Elicit hackathon attendee feedback regarding (via short, anonymous online survey):
    1. The three approaches for generating network representations
    2. The usefulness of network representations at different levels of annotation
    3. The value of inference methods for the different network representations

Inference of the generated networks may not be within the scope of the hackathon, but will be pursued after the meeting. During the hackathon, all code and generated network data will be developed for compatibility with existing BLAH tools and corpora and will be made available to the BLAH community. Finally, we will invite all interested community members to be authors in a manuscript drafted as a result of the proposed work.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors would like to acknowledge Drs. Kevin Cohen and Elizabeth White for their feedback and advice when drafting this proposal.

References

1. Albert R. *Network inference, analysis, and modeling in systems biology*. *Plant Cell*. 19(11):3327-38 (2007).
2. Piai S, Claps M. *Bigger Data for Better Healthcare. IDC Health Insights [Internet].*. *http://www.intelcore.cz/content/dam/www/public/us/en/documents/white-papers/bigger-data-better-healthcare-idc-insights-white-paper.pdf*. (2013).
3. Hunter L, Cohen KB. *Biomedical language processing: What's beyond PubMed?*. *Mol Cell*. 21(5):589-94 (2006).
4. Huang CC, Lu Z. *Community challenges in biomedical text mining over 10 years: success, failure and the future.*. *Brief Bioinform*. 17(1):132-44 (2016).
5. Neves M. *An analysis on the entity annotations in biological corpora. F1000Research [Internet].*. *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4168744/*. (2014).
6. Goldberg T, Vinchurkar S, Cejuela JM, Jensen LJ, Rost B. *Linked annotations: a middle ground for manual curation of biomedical databases and text corpora.*. *BMC Proc*. 9(5):A4 (2015).
7. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. *Cytoscape: a software environment for integrated models of biomolecular interaction networks.*. *Genome Res*. 13(11):2498-504 (2003).
8. Bernabò N, Barboni B, Maccarrone M. *The biological networks in studying cell signal transduction complexity: The examples of sperm capacitation and of endocannabinoid system.*. *Comput Struct Biotechnol J*. 11(18):11-21 (2014).
9. Lee M, Park K, Kim D. *Interaction network among functional drug groups.*. *BMC Syst Biol*. 7 Suppl 3:S4 (2013).
10. Bandyopadhyay A, Tsuji K, Cox K, Harfe BD, Rosen V, Tabin CJ. *Genetic analysis of the roles of BMP2, BMP4, and BMP7 in limb patterning and skeletogenesis*. *PLoS Genet*. 2(12) (2006).

Additional Files

There are no additional files.