Using Ontology Fingerprints to disambiguate gene name entities in the biomedical literature

Guocai Chen[1], Jieyi Zhao[1], Trevor Cohen[1], Cui Tao[1], Jingchun Sun[1], Hua Xu[1], Elmer V. Bernstam[1], Andrew Lawson[2], Jia Zeng[3], Amber M. Johnson[3], Vijaykumar Holla[3], Ann M. Bailey[3], Funda Meric-Bernstam[3], W. Jim Zheng[1,*]

[1]Center for Computational Biomedicine, School of Biomedical informatics, University of Texas Health Science Center at Houston

[2]Department of Public Health Science, Medical University of South Carolina, 135 Cannon Street, Suite 300, Charleston, South Carolina, 29425

[3]Department of Investigational Cancer Therapeutics, Institute for Personalized Cancer Therapy, UT-MD Anderson Cancer Center, 1400 Holcombe Blvd., FC8.3044, Houston, TX 77030

Personalized cancer therapy relies on extensive knowledge of cancer genes, their variants and treatments that target these variants. While most of this knowledge can be extracted from the biomedical literature, identifying genes and their associated publications with high precision is still a daunting task, often challenged by ambiguous gene names in the text. One way to disambiguate gene name is through gene normalization - the task of mapping a named entity in text to an identifier in a database. However, many genes have multiple names or aliases, part of them share identical names, even though they are distinct genes with different functions. Developing new methods to distinguish these ambiguous gene names will significantly improve the accuracy of information retrieval and other research-enabling applications.

To overcome this hurdle, we generated a non-supervised approach to create ontology profiles termed Ontology Fingerprints for selected genes that are relevant for personalized cancer therapy from the literature. The Ontology Fingerprint for a gene consists of a set of associated GO terms and their ancestors defined by biologists, with an enrichment p-value mapping to each term to reflect the significance of the term. We first used the ABGene/GNAT to identify gene names from the PubMed abstracts, and matched the names to the gene name or alias of known genes. The ambiguous names were then assessed by evaluating the degree to which the abstract matched the Ontology Fingerprints of the genes.

Focusing only on genes targeted by therapeutics for personalized cancer therapy. Eleven of these genes and relevant PubMed articles were selected and marked by oncologists and research staff from the Institute for Personalized Cancer Therapy at the UT MD Anderson Cancer Center. For the selected genes, we obtained 93.6% precision for gene name disambiguation and 80.4% AUC for gene and article association. For additional 223 human genes relevant to cancer, by using the Ontology Fingerprints generated from the publications before December 20, 2009 for these genes to predict the association of these genes with papers published after 2009, we got a highest precision up to 92.7%.

We investigated the feasibility of using Ontology Fingerprints to discover associations between genes and PubMed articles, as well as to disambiguate gene name mentions. We obtained reasonable accuracy for gene name disambiguation and gene and PubMed article association. The Ontology Fingerprint method can improve gene normalization and the analysis of gene and article association. We conclude that Ontology Fingerprints can help disambiguate gene names mentioned in text and analyze the association between genes and articles.

The core algorithm was implemented using a GPU-based MapReduce framework to handle big data and to improve performance. Comparing with running the program on Lonestar cluster, we can gain the same magnitude of speed when using the GPU MapReduce framework. Overall, the MapReduce framework makes execution of the program more convenient and affordable, especially on a workstation with an appropriate graphic card.

*Corresponding Author: Wenjin.j.Zheng@uth.tmc.edu