Evaluation of the Performance of BioNLP Tools for Discovering Causal Genes in Terms with Pathway Enrichment

Xuan Qin\textsuperscript{1,2}, Shuguang Wang\textsuperscript{2}, Yongze Wu\textsuperscript{3} and Jingbo Xia\textsuperscript{1,2,*}

\textsuperscript{1}College of Informatics, Huazhong Agricultural University, Wuhan, China; \textsuperscript{2}Hubei Key Laboratory of Agricultural Bioinformatics, Wuhan, China; \textsuperscript{3}College of Science, Huazhong Agricultural University, Wuhan, China; *Correspondence author: xajingbo.math@gmail.com, xjb@mail.hzau.edu.cn

Abstract. To meet the aim of large-scale knowledge discovery, biomedical natural language processing is regarded as an effective tool to curate hidden information in research papers or related texts. We proposed a new TF*IDF based metric, Gene Enrichment Metric, and carried on thorough comparison among p-value and F-score. Subsequently, we used this metric to evaluate the effectiveness of three baseline BioNLP tools in the pathway enrichment analysis. Actually, pathway enrichment analysis plays an important role in overlying functioning mechanism of disease and causal genes. Henceforth, the design of evaluation metric in terms with pathway enrichment shed light on the key knowledge discovery and its reproducibility in in-silico bioinformatics computation.

1. Introduction

Knowledge discovery based on text mining technique has played a key role in the knowledge curation and new relation extraction in the field of bio-medical research [1]. The innovative job was Swanson’s ABC model, where relationship between drugs and clinical purposes was mined out manually and evidence show that fish oil has new therapeutic effect upon Raynoud’s syndrome [2].

Till now, dozens of Biomedical natural language processing (BioNLP) tools have been developed to incorporate the usage of bio text data and ontologies, and various task settings are designed. Co-occurrence strategy by associating co-occurred entities has been extensively used in the context of literature-based discovery, so as to creating interconnection between genes, protein, and other biological entities. More frequently the co-occurrence shows, higher the chance is for the entity-pair relevance. However, this strategy relied heavily on the quality of NER (Named entity recognition) tools, and it took years to develop sophisticated tools like Pubtator [3], and TEES [4]. Here, pubtator is regarded as a standard bio-term NER tool, which with high accuracy annotates drug, protein, gene, mutation in scientific literature in its plain text form.

Though co-occurrence played vital roles in knowledge discovery, its results lack sufficient semantic evidence to support biomedical findings. Sophisticated semantic relationship miner and algorithms were developed to enhance the text mining methodologies. The Turku Event Extraction System [4] is the best performing tool in the GENIA BioNLP 2009/2011 shared tasks, which was designed for the extraction of events and relations from biomedical text. In its strategy, entities are annotated first and sentences are parsed, and then an event detector is used to detect trigger words and relationship type between entities. By doing this, TEES find relation between two entities which are not only co-occurrence, but also semantic-oriented.
The purpose of this research is to lead an evaluation for calculating the performance of text mining tools in a drug discovery scenario by using BioNLP strategy. The scenario assumes a text miner, with curated causal genes, accurately confirms a target pathway in which causal genes play vital roles. Thus, a drug discovery paradigm will be achieved accordingly. In this specific scenario, the evaluation of the selection of causal genes relies on both the amount and the extend of genes that cluster in a target disease pathway. For this purpose, we proposed a Gene Enrichment Metric (GEM), and compared it with traditional metrics like p-value in hypergeometric test, or F-score. Performance comparison is taken between two typical BioNLP tools, each of which represents typical task design in Bio text mining. Here, Pubtator for co-occurrence is designed for looking for rough knowledge representation of biomedical entities, and TEES is set for fine-grained knowledge representation.

2. Material and Method

2.1. Paradigm of Causal Gene Discovery by using BioNLP Tools or Strategies

As discussed in Introduction, one of the roles of BioNLP tool in bioinformatics knowledge discovery is to offer help for new-drug finders to curate causal genes that function as a cluster in a target pathway and bring vital effect for a target disease. This is a promising application in drug discovery from view of bioinformatics and chemoinformatics.

A common strategy of text mining tool in discovering disease-related cause gene and new drug is as the following:

- Step 1: Collect texts within the theme of a drug.
- Step 2: With computational NLP method, filter out key genes which are relevant with the drug.
- Step 3: Target a disease pathway, in which these genes play functions.
- Step 4: Collect external evidence to show the correlation of the drug and the disease/pathway.

However, there are considerations for the above strategy. First, each text mining tool has various drawbacks in performance, so a metric is need to evaluate different tools; Second, a metric is needed to evaluate the knowledge reproducibility ability of the tool, by assuming a hypothesis that causal genes cluster is enriched in a target pathway and it leads researcher to find the target disease pathway without prior knowledge; Third, generally, with an input genes list, there are dozens of pathway choices after enrichment analysis, and a metric is needed to evaluate the probability of top-ranking the target disease pathway.

2.2. BioNLP tools: Three Typical text Mining Tools: Pubtator (Abstract / Sentence Level) and TEES

Pubtator (Abstract Level) [3]: Pubtator (http://www.ncbi.nlm.nih.gov/bionlp/pubtator) is a PubMed-wide bio-entity recognition toolkit, which integrates a mount of entity recognition tools, including the dnorm and tmvar. High quality named entity annotations, including chemicals, mutations, diseases, genes and species are extracted from the medical literature. We collect all the gene entities via Pubtator. These genes are hopefully enriched in all drug-related abstracts, and are ranked by frequency, assuming that the genes on the top list are more active or causal than others.

Pubtator (Sentence Level) [3]: We collect the bio-entities that appear in the same sentence, and the co-occurrence in the sentence relationship is considered to be an interaction between genes. Genes that are more closely linked to other bio-entities are regarded as active and causal genes.

TEES [4]: The Turku Event Extraction System (TEES) is a typical fine-grained BioNLP tool, which was first proposed in the GENIA BioNLP 2009/2011 shared tasks contest. The algorithm of TEES was based on GENIA corpus for bio-curation. By using TEES, the trigger words can be extracted through the entity identification and the relationship between entities simultaneously. Genes co-occurred with trigger words were filtered out as causal genes after using TEES.

2.3. NLP task Setting: Pathway Enrichment

Pathway enrichment is set as a task here for performance evaluation of various BioNLP tools. A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in the cell. Pathways can trigger the assembly of new molecules, such as proteins, or turn genes on and off, etc. There are many types of biological pathways, the most common of which are...
pathways involved in metabolism, in the regulation of genes and in the transmission of signals. Since pathways play a key role in advanced research in genomics, pathway enrichment database has long been developed. Among all of the pathway database, Kyoto Encyclopedia of Genes and Genomes (KEGG) is a utility database dealing with genomes, biological pathways, diseases, drugs, and chemical substances, which also offers data analysis in genomics, metagenomics, metabolomics and other omics studies.

Three drug-pathway prior knowledge entries are used in the task testing step. First, Rapamycin has therapeutic effect on pancreatic cancer, and active genes clustered to the drug are expected to cluster in the pancreatic cancer pathway. Similarly, Avastin to colorectal cancer, and Lorsatan to Renin angiotensin system disorder, are set as drug-pathway pairs under consideration. Text data for fulfilling the pathway enrichment come from PubMed (https://www.ncbi.nlm.nih.gov/pubmed/), a huge repository of biomedical publications. Abstracts related to a disease topic are downloaded, and key gene entities are extracted from the above three BioNLP methods. The data details are listed in simulation test in Result.

From this task setting, text mining tools are evaluated by considering the effectiveness of retrieved causal genes and observing the density of corrected curated genes in the target pathway. Generally, the density of gene enrichment on the target pathway shows higher performance of the knowledge curator. In another word, if a big portion of the known causal gene were filtered out by a text mining tool and fell into target pathway, the tool is regarded as an effective tool in pathway enrichment. Henceforth, suitable metrics are necessary to evaluate the performance of various BioNLP methods in terms with causal gene finding and pathway enrichment.

2.4. Traditional Metrics for Pathway Enrichment or Information Retrieval

2.4.1 P-value: traditional metric for pathway enrichment. In tradition, the idea of p-value computation counts if it is needed to evaluate how exactly a cluster of genes is able to be enriched in a targeted pathway. Here, $P(x) = C_x^n \times C_{k-x}^{N-n} / C_k^N$. As a popular pathway enrichment tool in bioinformatics, The p-value used in clusterProfiler [5] is: $P = 1 - \sum_{j=0}^{x-1} p(i)=P(x) + P(x + 1) + \ldots + P(k)$. Here is an example to show that a simple comparison of p-values failed to make p a perfect metric. Here, genes clustered upon a drug, Auranofin, are enriched in two pathways, i.e., Influenza A and Tuberculosis. Since TEES only find six genes after text mining PubMed citations of the corresponding drug, the point here is that 3 or 4 out of 6 genes should not vary such huge in p-values, as shown in table 1. However, the p-value consideration upon the pathway enrichment has two flaws.

| Target pathway | Gene ratio x/k | Gene ratio m/N | p-value |
|----------------|----------------|----------------|---------|
| Influenza A    | 4/6            | 173/7404       | 4.16E-06|
| Tuberculosis   | 3/6            | 179/7404       | 2.634E-04|

The results are collected from TEES, which performed good in precision but rather low in recall rate. Hence there are only 6 out of 7404 genes are filtered. The computation is based on a hypothesis that the probability of having each gene is the same, which is not a reasonable prerequisite, e.g., house keeping genes generally participate in various pathways, but in the contrary some specific genes only appear in correlated pathway. Another impractical prerequisite is the requirement of the size of the gene set, k, to be a fixed value, since the size of the gene set varies if text mining tool changed.

2.4.2 F-score: traditional metric for information retrieval. In the statistical analysis of the binary classification, the F1 score (ie, the F-score or F-metric) is a classic and valid metric used to represent the accuracy of the classification. Here we can also simplify the KEGG pathway enrichment process to
a binary classification, and then use F1 score to compare the ability of different BioNLP tools to extract target pathway information. As an assessment indicator, the F1 score, usually stated as F-score, is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0. 

\[ F = \frac{2P \times R}{P + R} = \frac{2TP}{2TP + FP + FN} \]

Similar as p-value, F-score fails to represent the difference in gene functioning.

2.5. Proposed Metric for Evaluating the Effectiveness of Finding Causal Genes In Terms with Pathway Enrichment

As mentioned in the above, the metric for evaluating the pathway enrichment of a text-mining tool is to evaluate how good the system can curate specific genes for a correlated pathway. Then, the p-value metric need to be revised accordingly.

For a fixed pathway PW, text mining tool collects a gene set GS which contains k different genes, GenePW,1, GenePW,2, ..., GenePW,k. Label each gene with different TF*IDF score, 

\[ TFIDF(Gene_{PW,i}) = \text{term frequency}(Gene_{PW,i}) \times \text{inverse doc frequency}(PW) \]

Hence, the higher the value is, the more specific the gene clusters to the pathway. So we formulated a score to each gene, to quantify how relevant this gene is to a fixed pathway. So, for the pathway PW and the gene set GS, a gene TF*IDF sum (GTS) is obtained by summing the TFIDF score of the k genes:

\[ GTS(PW, GS) = \sum_{i=1}^{k} TFIDF(Gene_{PW,i}) \]

Based on the notations in equation, please note that there are only x out of k genes are actually related to pathway PK, and for the x–k miss-targeting genes, the TF*IDF scores equal to zero. For clarifying the idea, here comes an example of for the rapamycin-related gene enrichment in pancreatic pathways. In table 2, p-value comparison is shown in the fifth column. The results are collected from different Text Mining tools, i.e., Abstract level TM, Sentence level TM, and TEES TM. The three has descending order in recall, but ascending order in precision. From ClusterProfiler p-value, the Sentence level TM has greatest significance in gene enrichment if compared with Abstract Level TM and TEES TM, while TEES performed fairly awkwardly.

Table 2. An example of p-value comparison for the rapamycin-related gene enrichment in pancreatic pathways.

| Target pathway | TM method | Gene ratio x/k | Gene ratio m/N | p-value |
|----------------|-----------|----------------|----------------|---------|
| Pancreatic cancer | Abstract | 66/3905 | 75/7404 | 3.16E-34 |
|                 | Sentence | 57/1514 | 75/7404 | 3.05E-35 |
|                 | TEES | 13/290 | 75/7404 | 3.87E-11 |

To introduce the metric for gene enrichment for GS, the GTS (PWPancreatic, GSAbstract), GTS(PWPancreatic, GSSentence), GTS(PWPancreatic, GSTEES) are added in the sixth column.

We want a metric, which (1) doesn’t rely on k, and (2) it represents the GTS value of a GS, and (3) it as well reflects the size of GZ, so as to quantify how good the gene set is significant enriched to a fixed pathway. Therefore, a gene enrichment metrics (GEM) is proposed here:

\[ GEM(PW, GS, k) = GEM(k) = \frac{GTS(PW, GS)}{k} \]

For clarifying the use of GEM metric, we assume there is a novice who claims that all the 7404 genes are what he likes to get. In addition, there is an expert text mining tool, which collects 76 genes, and 75 correlated genes out of 76 genes are obtained. The result in Table 3 show sufficient reasonability. The novice and expert obtained the lowest and high score, while TEES and Pubtator (Sentence) outperform Pubtator (Abstract level).
Table 3. An example of p-value comparison for the Rapamycin-related gene enrichment in Pancreatic pathways.

| Target pathway   | TM method | Gene ratio x/k | Gene ratio m/N | p-value | GTS   | GEM |
|------------------|-----------|----------------|----------------|---------|-------|-----|
| Pancreatic cancer| Abs.      | 66/1639        | 75/7404        | 3.16E-34| 4.77  | 0.00122 |
|                  | Sent.     | 57/989         | 75/7404        | 3.05E-35| 4.10  | 0.0027 |
|                  | TEES      | 13/108         | 75/7404        | 3.87E-11| 0.93  | 0.0032 |
|                  | Novice    | 75/7404        | 75/7404        | NA      | ~ 5   | 0.0006 |
|                  | Expert    | 75/7404        | 75/7404        | NA      | ~ 5   | 0.0526 |

3. Result

Rapamycin, a drug for pancreatic cancer, Avastin, a drug for Colorectal cancer and Losartan a drug target Reninangiotensin system are part of 22 target drugs for this test. 31, 118 abstracts related to rapamycin, 13501 abstracts related to Avastin and 8,837 abstracts related to Losartan in PubMed (US National Library of Medicine National Institutes of Health) are retrieved as the original text data. All abstracts dealt with three text mining method namely TEES, co-occurrence-sentence (Sent.) and co-occurrence-abstract (Abs.). Three different active gene entity groups recognized by three text mining method are enriched by KEGG through R package clusterProfiler. The statistical result data are shown in Table 4.

Table 4. Statistical results of three text mining methods about three cases.

| Method   | Pathwayrank | Gene ratio x/k | Gene ratio m/N | p-value | GTS   | GEM | F-score |
|----------|-------------|----------------|----------------|---------|-------|-----|---------|
| Case 1:  | Rapamycin   | Abstract 3     | 66/1639        | 75/7404 | 3.16E-34| 4.77 | 0.003 | 0.077  |
|          |             | Sentence 2     | 57/989         | 75/7404 | 3.05E-35| 4.10 | 0.004 | 0.107  |
|          |             | TEES 7         | 13/108         | 75/7404 | 3.87E-11| 0.932| 0.009 | 0.142  |
| Case 2:  | Avastin     | Abstract 25    | 33/499         | 86/7404 | 1.82E-17| 2.030| 0.004 | 0.113  |
|          |             | Sentence 251   | 17/222         | 86/7404 | 4.42E-10| 1.014| 0.005 | 0.110  |
|          |             | TEES 40        | 5/43           | 86/7404 | 1.28E-04| 0.302| 0.007 | 0.078  |
| Case 3:  | Losartan    | Abstract 76    | 9/383          | 23/7404 | 1.03E-06| 2.762| 0.007 | 0.044  |
|          |             | Sentence 47    | 7/208          | 23/7404 | 2.09E-06| 2.017| 0.010 | 0.061  |
|          |             | TEES 57        | 5/108          | 23/7404 | 1.64E-05| 1.321| 0.012 | 0.076  |

Case 1: Drug Rapamycin active genes enriched to pathway Pancreatic cancer. Case 2: Drug Avastin active genes enriched to pathway Colorectal cancer; Case3: Drug Losartan active genes enriched to pathway Reninangiotensin system disorder

The simulation results showed the idea of paradigm which is introduced in Section 2.1. For example, in case 1, there is prior knowledge showing that 75 causal genes exist in pancreatic pathway, and the BioNLP tool –Pubtator-Abstract– extract 1,639 genes out of 7,404 background gene set. Among the filtered genes, 66 ones are causal genes. The p-value of the pathway enrichment is 3.16E-34, which shows high significance in stating that the genes are successfully enriched in the pancreatic cancer pathway. This p-value itself support the fact that the causal genes related to the drug Rapamycin are correlated to the target pathway. Being similarly in case 1, the other two BioNLP tools, i.e., Pubtator-abstract and TEES, also show nice performance in causal gene filtering. For Pubtator-abstract, 57 genes out of 989 filtered genes are causal genes, while for TEES; the ratio is 13/108. Both tools achieved sufficient significance, as the p-value is 3.05E-35 and 3.87E-11, respectively.

Though the simple analysis showed that all of three BioNLP tools performed well in pathway enrichment, the purpose of the simulation test is to show the consistency in terms with various metrics.
As shown in the fifth column, when discussing \( p \)-value, the three BioNLP tools performed with different manner in three cases. In case 1, Pubtator-Abstract was the best, and TEES performed poorly. In case 2, Pubtator-Abstract outperforms the others. While in case 3, Pubtator-Sentence is the best. As mentioned before, \( p \)-value is not a suitable metric for causal gene evaluation, and it failed to show consistency in the pathway enrichment simulation test. Similarly, F-score failed to obtain consistency to distinguish the advantage among the three methods in three cases.

Instead, from GEM metric, a consistent result was obtained, and it showed that TEES is regarded as the one with best performance in pathway enrichment. Here, TEES obtained the highest GEM score in case 1, 0.009, and the scores of the other two are 0.003 and 0.004 respectively. In case 2, the GEM score for TEES, Pubtator-Abstract and Pubtator-Sentence are 0.007, 0.005, and 0.004, respectively. In case 3, the scores are 0.012, 0.010, and 0.007, separately. This high consistency showed that GEM is a stable performed metric, as well; it also proved that TEES is a better BioNLP tool to curate causal genes in terms with pathway enrichment.

4. Discussion
The discussion in results show that neither \( p \)-value nor F-score show consistent performances for three testing cases, and that make GEM the only alternative for causal gene clustering evaluation. As discussed in Table 4, \( p \)-value is a good metric for traditional pathway analysis, but fails as it relies on a fact that each gene has the same chance to be clustered into a specific pathway. However, this is not the case in biology. Because of the same reason, F-score is not suitable for the causal gene clustering scenario, though it is a good metric for information retrieval. In GEM, TF * IDF calculation reflects the prior distribution of different genes in various pathway, and make it possible to calculate a metric which relies both on pathway knowledge and gene clusters. Second, GTS is an accumulative effect of TF * IDF and it reflects the performance of gene set as a cluster. Eventually, average GTS, i.e., GEM, leverage the size and the density of clustering accuracy. With GEM, text miners are hopefully able to find potential text mining tools which is more capable of finding causal genes and mapping them to a target disease. Hence the results are collected in Table 5.

### Table 5. Comparison of metrics.

| Evaluation                              | \( p \)-value | F-Score | GEM  |
|-----------------------------------------|---------------|---------|------|
| Good metric for pathway enrichment      | Yes           | -       | -    |
| Good metric for information retrieval   | -             | Yes     | -    |
| Good metric for causal gene clustering  | No            | No      | Yes  |

5. Acknowledgement
This work is funded by the Fundamental Research Funds for the Central Universities of China (Project No. 2662018PY096). We expressed our gratitude to Kevin B Cohen and Jari Bjorne for the discussion of knowledge granule and reproducibility.

6. References
[1] Verspoor, Karin, et al. "A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools." BMC bioinformatics 13.1 (2012): 207.
[2] Swanson, Don R. "Medical literature as a potential source of new knowledge." Bulletin of the Medical Library Association 78.1 (1990): 29.
[3] Wei, Chih-Hsuan, Hung-Yu Kao, and Zhiyong Lu. "PubTator: a web-based text mining tool for assisting biocuration." Nucleic acids research 41. W1 (2013): W518-W522.
[4] Björne, Jari, et al. "Drug-drug interaction extraction from biomedical texts with svm and rls classifiers." Proceedings of DDIExtraction-2011 challenge task (2011): 35-42.
[5] Yu, Guangchuang. Using clusterProfiler to identify and compare functional profiles of gene lists. (2013).