Context-aware query design combines knowledge and data for efficient reading and reasoning

Emilee Holtzapple[1†], Brent Cochran[2‡], Natasa Miskov-Zivanov[1,3,4†]
[1]Dept. of Computational and Systems Biology, [2]Dept. of Developmental, Molecular, and Chemical Biology, [3]Dept. of Electrical and Computer Engineering, [4]Dept. of Bioengineering
[†]University of Pittsburgh, Pittsburgh, PA USA.
[‡]Tufts University School of Medicine, Boston, MA USA.
erh87@pitt.edu, Brent.Cochran@tufts.edu, nmzivanov@pitt.edu

Abstract

The amount of biomedical literature has vastly increased over the past few decades. As a result, the sheer quantity of accessible information is overwhelming, and complicates manual information retrieval. Automated methods seek to speed up information retrieval from biomedical literature. However, such automated methods are still too time-intensive to survey all existing biomedical literature. We present a methodology for automatically generating literature queries that select relevant papers based on biological data. By using differentially expressed genes to inform our literature searches, we focus information extraction on mechanistic signaling details that are crucial for the disease or context of interest.

1 Introduction

The number of peer-reviewed publications in molecular biology, biotechnology, and biomedical research increases exponentially every year. There is a considerable number of published papers on any one mainstream biomedical research topic, potentially hundreds of thousands of relevant articles. For many areas of study, simply reading every paper is unrealistic, or even physically impossible. When studying biological systems, such as intracellular signaling networks, this problem is apparent – accurate representation of all relevant signaling events requires extensive, expert knowledge acquired over many years of study. By using natural language processing, machine readers are capable of extracting interactions from hundreds or thousands of papers in a matter of hours, achieving a substantial speedup over manual information extraction (Björne & Salakoski, 2011). For this reason, automated methods for information extraction, such as machine reading, are used to retrieve information about intracellular signaling networks, and this information can then be used for model assembly or extension. While automated methods accelerate model assembly, the time required for processing all selected papers still depends on the number and the type of papers chosen for machine reading (Holtzapple, Telmer, & Miskov-Zivanov, 2020).

To retrieve relevant papers for machine reading, a common method is to query databases that contain biomedical literature. One repository for biomedical literature, MEDLINE, contains over 27 million papers (Fiorini, Lipman, & Lu, 2017), and a common method for retrieving papers from MEDLINE is through its associated search engine, PubMed. Querying MEDLINE through PubMed is particularly useful for identifying papers on a specific context such as disease or cell type. It is also used for identification of individual proteins, signaling pathways, and general cell processes in one specific context. One example of a PubMed query that targets a single pathway in a specific context is ‘"Hippo pathway" AND "stem cells"’. This query returns 272 papers, many of which describe Hippo pathway signaling trends in cancerous stem cells, as well as non-cancerous stem cells. These papers contain a wealth of information about the mechanistic causes of stemness. However, retrieval of these papers requires a priori knowledge that the Hippo pathway is important in stem cell maintenance and renewal. Additionally, these papers describe one small facet of stem cell signaling, and do not contain all the information needed to understand the system. To widen our perspective, we could retrieve all papers in MEDLINE that concern stem cells by querying PubMed with "stem cells". Here, we encounter two obstacles – this query returns over 271,000 papers, many of which describe morphological or anatomical details, and not signaling pathways.

Proceedings of the BioNLP 2021 workshop, pages 238–246

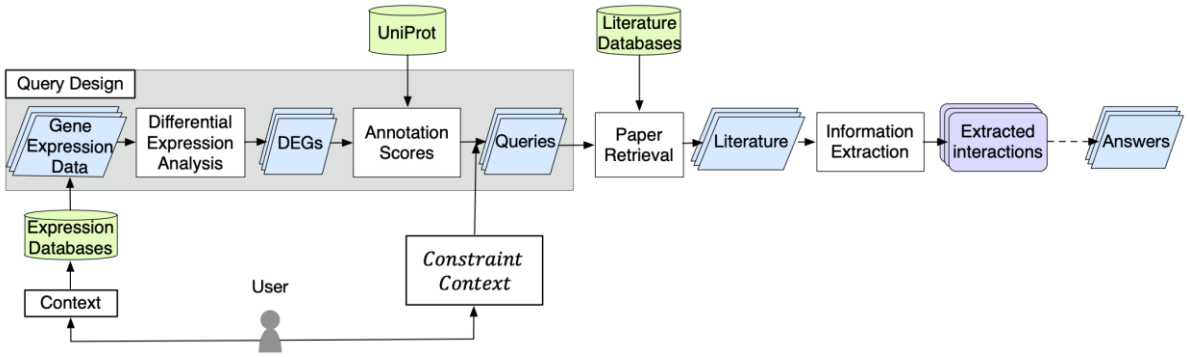


Figure 1. The automated query design methodology for information retrieval in biomedical research.

Our dilemma is that retrieval of relevant, crucial papers requires prior knowledge of which pathways are important.

There is a demand for improvement in methods for patient-specific paper retrieval, as evidenced by the TREC precision medicine track (Roberts et al., 2017). State-of-the-art methods for paper retrieval rely on term lists generated by experts or users, or automated information retrieval of similar papers (Sesagiri Raamkumar, Foo, & Pang, 2017; Wesley-Smith & West, 2016). These methods have several disadvantages. First, paper retrieval may depend on the cooperation of one or more experts in the field. Even for automated techniques that locate papers through related citations, or sematic analysis, some level of prior knowledge is needed. Also, even for experts that are up to date on canonical signaling pathways in a context of interest, novel pathways or signaling events cannot be easily targeted in a literature query. For efficient, thorough, context-aware exploration of cellular signaling, improved methods for literature retrieval are needed.

We present here a methodology for automated query design that does not rely on manual steps of the domain expert. To address the potential role of differentially expressed genes (DEGs) in disease mechanisms, we infer queries from biological data. Under- or over-expressed genes in the disease state are often the ones that play a role in disease progression (Armstrong et al., 2002). Identifying these contextual DEGs and using them as query terms focuses literature reading on genes and proteins that have altered signaling trends, and therefore, it facilitates further exploration of intracellular signaling networks that are potentially affected in disease. Our method utilizes gene expression data to find possible genes of interest based on their relative expression changes in response to disease, infection, etc. These genes of interest are used in the query to narrow down all possible PubMed hits to relevant signaling papers only. Furthermore, we also take into consideration how well-known each gene is, to choose the optimal number of gene terms in a query. Our results show that automated query design using these methods returns relevant signaling papers, and interactions extracted from these papers are informative and useful when reasoning about the queried context. This addresses a well-established problem in precision medicine – altered signaling pathways are often unique to one patient or environment and are difficult to study manually. Our methodology can be used in conjunction with any state-of-the-art model assembly techniques to aid in understanding affected signaling mechanisms in patient or cell line-specific systems. This methodology will provide an automated framework to retrieve research papers and streamline the process of model assembly. Our proposed automated query design methodology is outlined in Figure 1.

2 Query Design Method

In the following sub-sections, we describe our method to identify DEGs in the context of a disease, cell line, tissue type, or other condition (e.g., drug treatments), and for using them to form query terms when searching literature.

2.1 Identification of differentially expressed genes

As shown in Figure 1, the first step in our query design method is to define a context for literature search. Our approach allows a user to automatically design queries for many different contexts, including any biological condition that can be observed long enough to generate gene expression

data. The user selects a data source and a relevant dataset from that source. While any kind of gene expression data can be used (microarray, RNA-seq, or single cell RNA-seq), public databases for expression data most frequently include RNA-seq data. Public databases for RNA-seq data include the Cancer Genome Atlas (TCGA) (Weinstein et al., 2013), Gene Expression Omnibus (Clough & Barrett, 2016), and the Expression Atlas (Papatheodorou et al., 2018), all of which contain sufficient expression data to be used in our proposed query generation method.

Once the dataset file is selected and input by the user, our proposed query design method identifies genes that are differentially expressed in the context of interest (e.g., disease state, cell line, etc.), compared to the control. The RNA-seq technique provides insight into the transcriptional activity of a cell population and reveals the number of gene transcripts present at a single point in time. For any gene *X*, we compute its differential expression as the fold change between the amount of its transcript ($X_{transcript}$) in two scenarios, control ($X_{transcript}^{control}$) and disease state ($X_{transcript}^{disease}$), a common method for measuring changes in gene expression (Huang, Zhang, Shen, Wong, & Xie, 2015). Since in this work we are interested in the magnitude of the change from the control, and not the direction of the change (i.e., increase or decrease), we use the absolute value of the change:

$$d_X = \left| \frac{X_{transcript}^{disease}}{X_{transcript}^{control}} \right| \quad (1)$$

We determine the d_X value for all transcripts in the selected RNA-seq dataset. Next, we sort the transcripts in a descending order of their d_X values (i.e., descending magnitude of change), and select a threshold for the d_X value, to ensure that all genes used as query terms are relevant to the dataset context. Specifically, we use 2.0 as a threshold, that is, we remove from the sorted list those transcripts that have $d_X < 2.0$. The standard threshold for d_X is usually 2.0 or 1.5 (Huang et al., 2015), based on what a cell biologist would consider notable or likely due to the effect of the disease or altered state, and not just noise in gene expression. While we use $d_X \geq 2.0$, the user can adjust this threshold to suit the research context (i.e., diseases or cell types with more or less DEGs than expected). We will refer to the transcripts remaining in the sorted list as DEGs. As probable indicators of a disease state, these DEGs become candidates for query terms. To give an estimate of an expected size of the sorted DEG list, previous work on analyzing many RNA-seq datasets over a wide range of conditions, including disease, tissues, cell types, drug treatments, etc., has shown that the median number of DEGs (with $d_X \geq 2.0$) per dataset is 92 (Crow, Lim, Ballouz, Pavlidis, & Gillis, 2019). However, as many as 10,000 DEGs per dataset were also observed, although rarely. We expect to see dozens to hundreds of DEGs (gene transcripts with $d_X \geq 2.0$), out of the 20,000+ genes in an RNA-seq dataset.

2.2 Selection of query terms

The sorted list of selected context-dependent DEGs that is automatically generated as described in Section 2.1, and the list of context terms, $Context$, provided by a user, are inputs to the next step of our proposed query generation method.

Using all DEGs with $d_X > 2.0$ to formulate a query is still not practical, as there can be tens or hundreds of such DEGs (see Section 2.1). Instead, we propose a method to further reduce the size of the sorted DEG list. We determine the number of DEGs to be used as query terms by estimating the number of papers that would be retrieved from a literature database when using the query formed from these terms. For example, in PubMed, the "popularity" of genes varies widely: *TP53* is a well-known oncogene with over 100,000 papers found in PubMed, and therefore, any query containing "p53" will return more papers than a query using a novel gene.

Thus, to estimate the impact of each DEG, as a possible query term, on the number of papers retrieved, we propose to utilize the annotation information provided by the UniProt database (The UniProt Consortium, 2017). This database contains information on the gene itself, known transcripts, as well as information on the gene product, if available. Each gene in the UniProt database has an assigned *annotation score*, which is an amalgamation of evidence of the gene and gene product's existence, including cross-references in other databases, known aliases, experimental evidence, and more. We use this annotation score as a measure of how established a gene is in the literature. For manually annotated genes, where the evidence has been reviewed by an expert, the score is higher. The annotation score has

an integer value in the interval between 1 and 5, where score of 5 indicates ample evidence of the protein in existing literature and databases, and score of 1 indicates little to no available information about the protein. For example, the *TP53* gene in humans (UniProt ID P04637), a well-known tumor suppressor, has an annotation score of 5, while the *OATL1* transcript in humans (UniProt ID B4DF03), which has not been observed at the protein level, has an annotation score of 1.

We propose here to use the annotation score together with the d_X value when deciding which DEGs to include as query terms. The combination of these two measures allows the design of queries for different objectives or tasks, for example, to search for literature that contains a few well-known (high annotation score) proteins, or many novel or unstudied (low annotation score) proteins. Furthermore, by incorporating the UniProt annotation score to choose terms, we can automatically design queries that will lead to a selection of a manageable number of papers. In other words, the optimal number of papers would be the one large enough to provide adequate information on the system and small enough to still be processed in a feasible amount of time. What would be considered the "optimal" number of papers depends on both the complexity of the context of interest, as well as the allocated resources for information extraction. For example, a researcher using a machine reader to process literature on diabetes will require many more papers than someone who wants to read papers manually. Additionally, the number of papers found in a literature database as a result of the query will be different for each user depending on the input dataset, annotation score, and the addition of new publications in the literature database, and so this method allows to tailor the query design process to the user's research goals. We will refer to the DEGs that are selected to be used in a query as *query term DEGs*.

Different research tasks, paper contexts, and datasets will require a different number of papers to be read. Therefore, our method allows the user to provide an additional input, $Constraint$, which will influence the number of papers selected for reading. The $Constraint$ input can be either categorical, or a discrete number greater than 0, and is used in our method to determine the cut-off parameter, C. The cut-off C value is in turn used to select those DEGs that will be included in the query. Specifically, we traverse the sorted DEG list, starting with the DEG that has the largest d_X value, and we keep adding DEGs to the query term list, as long as the sum of their annotation scores is smaller than or equal to the cut-off value C.

Table 1. User-input categories, the corresponding cut-off parameter C for the annotation score sum, as well as the expected maximum and minimum number of query term DEGs. (These values do not account for DEGs with no entry in the UniProt database.)

user-input category	C	expected min # of DEGs	expected max # of DEGs
human-readable	15	3	15
automation suggested	35	12	35
automation required	60	20	60

We use three categories to indicate the level of automated reading needed to comprehend all information in the paper set. The first category, "human-readable", results in a selection of a small number of papers, suitable for a human to read in a short time (e.g., hours). The second category, "automation suggested", leads to a medium number of selected papers that is possible for a human to read (e.g., days), but more practical if processed by machine reading. The third category, "automation required", results in a large number of selected papers, only practical for machine reading.

Allowing for two different ways to enter the $Constraint$ input, provides additional flexibility. If the user knows exactly which value they want to use for the cut-off parameter, they can directly enter it. However, in the research process, the users may sometimes be interested in exploring a smaller subset of relevant papers, or doing a more comprehensive exploration of the topic, and the three categories listed above are useful in such cases. The values of the parameter C that correspond to the three categories, and that we used to obtain results and demonstrate our approach, are listed in Table 1.

We note here that, while these values are set internally in the code, they could be easily changed to better suit different domains or research goals. For example, for a "human-readable" reading output, we set C=15, and following our method for selecting query term DEGs given the cut-off value C, this could result in as few as 3 query term DEGs (all with annotation score 5) or as many as 15 query term DEGs (all with annotation score 1).

To this end, it is worth noting that not all DEGs are always found in UniProt, and therefore, the DEGs without a corresponding UniProt entry are assumed to have annotation score value of 0. As this is possible even for DEGs with large d_X value, this could lead, in rare cases, to the actual number of query term DEGs exceeding the cut-off value C (e.g., this would be 15, for our example above). While, in theory, the number of DEGs with $d_X \geq 2.0$ and annotations score of 0 could potentially be very large, we have not encountered such cases. Moreover, our experiments have shown that allowing for DEGs with annotation score 0 to be added to the query term list does not significantly increase the number of selected papers, while at the same time can lead to the retrieval of papers with very novel disease mechanisms. In Table 1, we provide the C values that we use for the three user-input categories, and the corresponding typical minimum and maximum number of query term DEGs. As a guidance, we list in Table 1 the *typical* minimum and maximum numbers that are easily determined from C values, and which consider only those genes with an annotation score greater than 0.

Once the list of the query term DEGs is determined, their official gene names (e.g., *TP53, BRCA1, EGFR*) are combined with a logical **OR**, thus allowing any paper that includes at least one of the query term DEGs to be selected. It is important to note that the official gene name (or another standardized identifier) is already supplied by gene expression datasets, and so we avoid the challenge of using named-entity normalization to automatically standardize the names of DEG query terms. We chose to use a logical **OR** to retrieve the maximum number of relevant papers for each query, since a logical **AND** would make the query more specific, and so restrict the number of papers. Other combinations of logical **AND** and **OR** between the terms in the query are possible and could be informed by the user or inferred if relevant information is available. This is beyond the scope of the work presented here and is one of the next steps that we plan to explore in the future.

Furthermore, since we are interested in creating queries that focus on a particular context, our automated tool adds *Context* to this logical expression as a necessary condition, that is, it combines it with the other terms using a logical **AND**:

$$(gene_1 \textbf{ OR } gene_2 \textbf{ OR } \ldots\ gene_N) \textbf{ AND } Context \quad (2)$$

where each $gene_i$ (i=1,..,N) is the official gene name of one of the N query term DEGs. By including only papers that mention the context of interest, we can extract relevant interactions. It is important to note that one context may have multiple aliases (e.g., "coronavirus", "COVID-19", and "SARS-CoV-2" are all referring to the same disease). The user can increase the scope of the retrieved papers by combining all possible context aliases with a logical **OR**.

2.3 Using queries in disease explanation

We discuss in this section the use of automatically generated targeted queries in information extraction conducted by machine readers, followed by automated reasoning about affected signaling networks and biological processes. For each query, we retrieve all machine reading statements in the INDRA database (Gyori et al., 2017) that are associated with at least one paper in our reading set. The INDRA database is a system that incorporates natural language processing tools and standardized databases to collect biomedical signaling events. INDRA relies on several different machine readers to process papers and supply information on signaling events. The interactions output by readers are directed, and therefore, they can be used in the process of assembly or extension of dynamic models, in order to explain mechanisms and timing of the disease. Although the query term DEGs that were selected following our method described in Sections 2.1 and 2.2 are likely to participate in these interactions, it is important to note that the interactions output by readers will include many other relevant genes and proteins. Thus, these extracted interactions are expected to provide the information on intracellular signaling networks that is potentially critical for the context originally selected by the user and included as a term in the generated query (equation 2).

To evaluate the relevance of extracted interactions, we assess what types of biological processes and signaling pathways these interactions are involved in. We use PANTHER (Mi, Muruganujan, Ebert, Huang, & Thomas, 2018) to calculate enriched Gene Ontology (GO) terms (Ashburner et al., 2000) in the protein-protein interactions within our interaction sets for each query. In the GO database, genes and proteins are annotated with known cellular functions. Each

Table 2. Eight automatically formulated queries for four diseases. Each disease has two associated queries, which are expected to retrieve different sized reading sets.

Context	# of DEGs with $d_X \geq 2.0$	C	Query
Thyroid carcinoma	5026	15	**Q1**: (GABRB2 or LIPH or KLHDC8A or LIPI) and "thyroid carcinoma"
		60	**Q2**: (GABRB2 or LIPH or KLHDC8A or LIPI or LINC02471 or PRR15 or MTRNR2L12 or CIDEC or RTL4 or SLIT1 or ZCCHC12 or TRPC5 or LRP4 or RXRG or METTL7B or CDH3) and "thyroid carcinoma"
Ulcerative colitis	1476	15	**Q3**: (AL035661.1 or HP or NECAB1 or MCEMP1) and "ulcerative colitis"
		60	**Q4**: (AL035661.1 or HP or NECAB1 or MCEMP1 or ANKRD22 or ARG1 or BMX or MMP9 or S100A12 or SCART1 or SLC2A14 or SLC1A3 or SLC12A5-AS1 or OLAH or ACHE) and "ulcerative colitis"
COVID-19	32	15	**Q5**: (MX1 or DDX60 or PARP9) and (SARS-CoV2 or COVID-19 or coronavirus)
		60	**Q6**: (MX1 or DDX60 or PARP9 or DDX58 or HELZ2 or CMPK2 or OAS3 or STAT1 or HERC6 or DTX3L or IFIT1 or SAMD9 or AL445490.1 or TAS2R4 or AC147651.1 or AC004253.1) and (SARS-CoV2 or COVID-19 or coronavirus)
Glioblastoma	3300	15	**Q7**: (HOXD9 or PLA2G2A or HOXD10) and (Glioblastoma or GBM)
		60	**Q8**: (HOXD9 or PLA2G2A or HOXD10 or HOXD13 or HOXA5 or HOXD8 or DLGAP5 or HOXA10 or SAA1 or HOXC10 or AC092017.1 or AC011742.1 or AL160286.1 or MIR663AHG or ELL2P1 or TOP2A or IGHA1) and (Glioblastoma or GBM)

GO term has a list of proteins involved in the biological process, and PANTHER calculates representation of all known GO terms for each interaction set. For GO terms that have a greater number of genes found in the interaction set than would be expected by chance, we consider this GO term statistically enriched. To assess whether enriched GO terms are similar, we use NaviGo to calculate the Resnik similarity score between all GO terms (described in (Wei, Khan, Ding, Yerneni, & Kihara, 2017)). By determining highly enriched GO terms, we can draw conclusions about what signaling pathways and biological processes are represented in our paper sets for each query.

3 Results

To demonstrate the usefulness of our automated query design methodology, we show results for four different contexts. For each context, we automatically design two queries, one with an expected large number of output papers, and one with an expected small number of output papers. These results illustrate how DEGs can be used to formulate queries that output relevant papers, and how the annotation score affects the volume of papers. We also show that the papers contain interactions that are closely related and are involved in the same GO biological processes.

3.1 Case studies

Using the Expression Atlas (Papatheodorou et al., 2018), we selected four publicly available RNA-seq datasets. These four datasets provide gene expression data for both control and disease state in SARS-CoV-2 (Blanco-Melo et al., 2020), ulcerative colitis (Mo et al., 2018), glioblastoma multiforme (Gill et al., 2014), and thyroid carcinoma (Costa et al., 2015). All four datasets express transcription in transcripts per million (TPM) and include the d_X values computed for the disease state with respect to the control state. In the following studies, we use the d_X values that are provided with selected datasets. With these case studies, we cover three substantial topics in biomedical research – autoimmune disorders, cancer, and viral infections. These diseases differ in the number of expected publications, largely due to the awareness of the disease itself. We chose several well-studied diseases, as well as several relatively unknown diseases as case studies, to show the utility of our methodology, regardless of the recognition of the system at hand. Using differential gene expression data from these diseases, we illustrate how biological data can provide valuable information for automatically designed targeted queries.

3.2 Selection of queries

To design a query that retrieves a small reading set, as discussed in Section 2.2, we explored the effect of the cut-off value C=15 for the annotation score sum, and to design a query that retrieves a large reading set, we use the cut-off value C=60. The queries generated for all four contexts for these two cut-off values are listed in Table 2. Notably, the same cut-off value C for different datasets may result in queries with a different number of terms. This can be explained by the UniProt annotation score of the top (with large d_X) DEGs in the datasets. Due to differences in experiment techniques, environmental conditions, or other

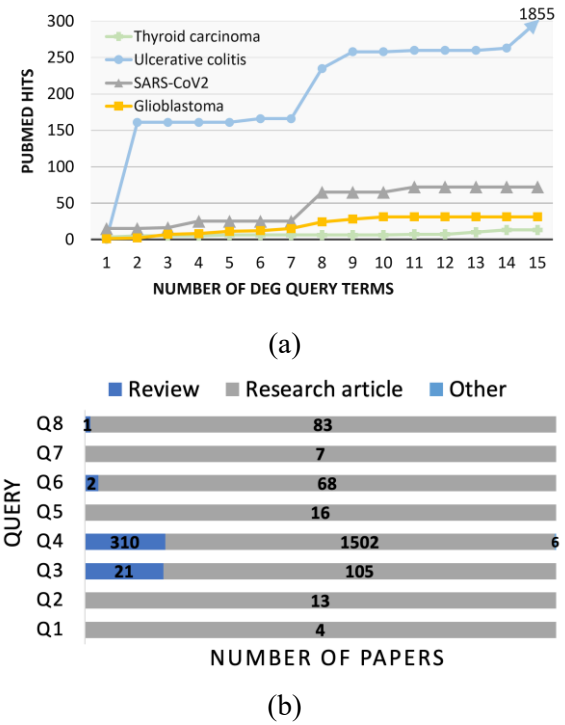


Figure 2. Number of papers found in PubMed, based on how many of the top DEGs were used as query terms. (b) Distribution of paper types by query.

factors, gene expression datasets from different samples and labs will likely show differences in the top DEGs. Consider a hypothetical example where we formulate queries based on two pancreatic cancer datasets (another example, not listed in Table 2), A and B, and choose the cut-off C=10. For dataset A, this value is achieved after adding two DEG query terms, since the DEGs with highest d_X values are P53 and MDM2, which are both very well-known proteins with an annotation score of 5. For dataset B, the threshold is not passed until five DEG query terms are added. The top five most differentially expressed genes are small non-coding RNAs, which are generally poorly studied, and each has an annotation score of 2.

3.3 Paper retrieval

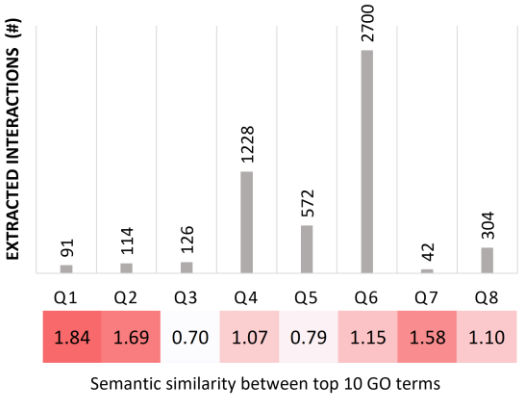
In our studies, we used PubMed (Fiorini et al., 2017) as the most up-to-date and comprehensive source for biomedical literature. We do not apply any filters for article type, year, or journal. However, we restrict our results to only those papers with valid PMCIDs, to ensure that all papers can be processed with state-of-the-art machine readers.

Once we have formulated queries for each disease, we can use them to search PubMed. In Figure 2a, we show the number of papers retrieved as a function of how many of the top DEGs are used as query terms. As expected, as the number of terms increase, so does the number of retrieved papers. However, many query terms, in conjunction with the context term, add no additional papers to the reading set. This indicates that some of these DEGs have not been explored much or mentioned in papers in the context of the relevant disease, and therefore, they may be a fruitful avenue for exploration.

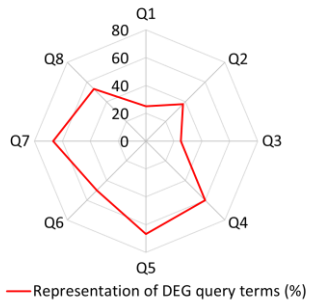
We also show in Figure 2b that, as the number of extracted papers in the reading output increases, the distribution of article types also changes. We examine the composition of the reading set by classifying each paper as either a research article, review, or other (books, documents, etc.). In large reading sets, reviews are slightly more common than in small reading sets, which is due to one or more query term DEGs having better representation in PubMed. Well-studied genes and proteins are more likely to be included in reviews than novel, relatively unknown genes. Since the scope of reviews and research articles differ drastically, we expect them to contribute differently to the number of extracted interactions.

3.4 Validation of extracted interactions

To validate the paper sets retrieved from each query, we analyzed the statements from the INDRA database (described in Section 2.3). In Figure 3a, we show the number of extracted interactions for each query. The number of interactions is dependent upon the number of papers, as well as the representation of the context and DEG query terms in PubMed. For each query, we also determined the top 10 enriched GO terms, sorted using the false discovery rate (FDR) (Benjamini & Hochberg, 1995). We show the average Resnik similarity score between the top 10 GO terms for each of our eight queries, where a higher score indicates more similarity between GO terms. Finally, in Figure 3b, we show the percent of DEG query terms that are present in the list of extracted interactions. These results, taken together, show that these queries retrieve papers that contain relevant signaling events that can be interpreted by machine readers, and describe highly related biological processes. In general, our method of increasing the cut-off value C not only retrieves



(a)



(b)

Figure 3 (a) Number of interactions extracted from INDRA for each query, as well as the average pairwise Resnik similarity score for the top 10 enriched GO terms. (b) The percent of DEGs used as query terms in each case study that are present in the set of extracted interactions.

more papers, but it also increases the number of signaling events extracted by readers, without a sizeable cost to relevance, as assessed by GO term semantic similarity.

4 Conclusions

While automated methods for extracting interactions from literature have improved the speed of information extraction, this process still has its pitfalls. Specifically, finding all relevant literature for the context at hand can be difficult, and brute force methods for selecting papers are too slow. By incorporating biological data in our queries, we can select relevant literature, and control the size of the reading sets.

Our results show that using DEGs to formulate queries allows for targeting literature that could help explain differentially regulated pathways in disease. One side effect of this method is identification of DEGs in disease where there is little to no literature presence. In such cases, our proposed method could become critical, as it automatically identifies the gaps in our collective knowledge of certain diseases, and thus, suggests important research directions. For DEGs that return no additional results when used as a query term, this indicates the gene has an undiscovered role in the context of interest.

Future directions include refining the query formulation methodology, as well as expanding our results. The relative presence of different diseases in PubMed affects the size of the reading set, independent of the number of gene query terms. By incorporating preliminary data on the presence of a disease or context in PubMed, we can adjust the annotation score. Additionally, since this method hinges on a list of affected genes or proteins with quantifiable differences from a control state, other measures of relative changes in cell function could also be used. Data on changes in post-translational modification of proteins, changes in epigenetic markers such as methylation, open chromatin, or histone modifications, or even somatic mutations could also be used, especially as such entities and events can be output by the state-of-the-art machine reading. Testing our methods on different datasets would help showcase the usefulness of our approach. In the future, we would also like to compare our method to a literature corpus assembled by an expert. However, since our queries are based on cell line-specific gene expression datasets, there are no existing corpuses for comparison. Future work includes assembling said corpuses and comparing to our method presented in this work.

Acknowledgements

Thanks to Yasmine Ahmed and Casey Hansen for helpful discussions. Funded by Defense Advanced Research Projects Agency (W911NF-17-1-0135).

References

Armstrong, Peter J, Johanning, Jason M, Calton Jr, William C, Delatore, Jason R, Franklin, David P, Han, David C, . . . Elmore, James R. (2002). Differential gene expression in human abdominal aorta: aneurysmal versus occlusive disease. *Journal of vascular surgery, 35*(2), 346-314.

Ashburner, Michael, Ball, Catherine A., Blake, Judith A., Botstein, David, Butler, Heather, Cherry, J. Michael, . . . Sherlock, Gavin. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics, 25*, 25. doi:10.1038/75556

Benjamini, Yoav, & Hochberg, Yosef. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological), 57*(1), 289-300.

Björne, Jari, & Salakoski, Tapio. (2011). *Generalizing biomedical event extraction*. Paper presented at the Proceedings of the BioNLP Shared Task 2011 Workshop, Portland, Oregon.

Blanco-Melo, Daniel, Nilsson-Payant, Benjamin E., Liu, Wen-Chun, Møller, Rasmus, Panis, Maryline, Sachs, David, . . . tenOever, Benjamin R. (2020). SARS-CoV-2 launches a unique transcriptional signature from in vitro, ex vivo, and in vivo systems. *bioRxiv*, 2020.2003.2024.004655. doi:10.1101/2020.03.24.004655

Clough, Emily, & Barrett, Tanya. (2016). The Gene Expression Omnibus Database. *Methods in molecular biology (Clifton, N.J.), 1418*, 93-110. doi:10.1007/978-1-4939-3578-9_5

Costa, Valerio, Esposito, Roberta, Ziviello, Carmela, Sepe, Romina, Bim, Larissa Valdemarin, Cacciola, Nunzio Antonio, . . . Ciccodicola, Alfredo. (2015). New somatic mutations and WNK1-B4GALNT3 gene fusion in papillary thyroid carcinoma. *Oncotarget, 6*(13), 11242-11251. doi:10.18632/oncotarget.3593

Crow, Megan, Lim, Nathaniel, Ballouz, Sara, Pavlidis, Paul, & Gillis, Jesse. (2019). Predictability of human differential gene expression. *Proceedings of the National Academy of Sciences, 116*(13), 6491. doi:10.1073/pnas.1802973116

Fiorini, Nicolas, Lipman, David J., & Lu, Zhiyong. (2017). Towards PubMed 2.0. *eLife, 6*, e28801. doi:10.7554/eLife.28801

Gill, Brian J., Pisapia, David J., Malone, Hani R., Goldstein, Hannah, Lei, Liang, Sonabend, Adam, . . . Canoll, Peter. (2014). MRI-localized biopsies reveal subtype-specific differences in molecular and cellular composition at the margins of glioblastoma. *Proceedings of the National Academy of Sciences, 111*(34), 12550-12555. doi:10.1073/pnas.1405839111

Gyori, Benjamin M., Bachman, John A., Subramanian, Kartik, Muhlich, Jeremy L., Galescu, Lucian, & Sorger, Peter K. (2017). From word models to executable models of signaling networks using automated assembly. *Molecular systems biology, 13*(11), 954-954. doi:10.15252/msb.20177651

Holtzapple, Emilee, Telmer, Cheryl A, & Miskov-Zivanov, Natasa. (2020). FLUTE: Fast and reliable knowledge retrieval from biomedical literature. *Database, 2020*. doi:10.1093/database/baaa056

Huang, H., Zhang, S., Shen, W. J., Wong, H. S., & Xie, D. (2015). Gene set enrichment ensemble using fold change data only. *J Biomed Inform, 57*, 189-203. doi:10.1016/j.jbi.2015.07.019

Maugeri-Saccà, M., & De Maria, R. (2018). The Hippo pathway in normal development and cancer. *Pharmacol Ther, 186*, 60-72. doi:10.1016/j.pharmthera.2017.12.011

Mi, Huaiyu, Muruganujan, Anushya, Ebert, Dustin, Huang, Xiaosong, & Thomas, Paul D. (2018). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research, 47*(D1), D419-D426. doi:10.1093/nar/gky1038

Mo, Angela, Marigorta, Urko, Arafat, Dalia, Chan, Lai, Ponder, Lori, Jang, Se Ryeong, . . . Gibson, Greg. (2018). Disease-specific regulation of gene expression in a comparative analysis of juvenile idiopathic arthritis and inflammatory bowel disease. *Genome Medicine, 10*. doi:10.1186/s13073-018-0558-x

Papatheodorou, Irene, Fonseca, Nuno A., Keays, Maria, Tang, Y. Amy, Barrera, Elisabet, Bazant, Wojciech, . . . Petryszak, Robert. (2018). Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Research, 46*(D1), D246-D251. doi:10.1093/nar/gkx1158

Park, J. H., Shin, J. E., & Park, H. W. (2018). The Role of Hippo Pathway in Cancer Stem Cell Biology. *Mol Cells, 41*(2), 83-92. doi:10.14348/molcells.2018.2242

Roberts, Kirk, Demner-Fushman, Dina, Voorhees, Ellen M., Hersh, William R., Bedrick, Steven, Lazar, Alexander J., & Pant, Shubham. (2017). Overview of the TREC 2017 Precision Medicine Track. *The ... text REtrieval conference : TREC. Text REtrieval Conference, 26*.

Sesagiri Raamkumar, Aravind, Foo, Schubert, & Pang, Natalie. (2017). Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems. *Information Processing & Management, 53*(3), 577-594. doi:https://doi.org/10.1016/j.ipm.2016.12.006

The UniProt Consortium. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research, 45*(D1), D158-D169. doi:10.1093/nar/gkw1099

Wei, Qing, Khan, Ishita K., Ding, Ziyun, Yerneni, Satwica, & Kihara, Daisuke. (2017). NaviGO: interactive tool for visualization and functional similarity and coherence analysis with gene ontology. *BMC Bioinformatics, 18*(1), 177. doi:10.1186/s12859-017-1600-5

Weinstein, John N, Collisson, Eric A, Mills, Gordon B, Shaw, Kenna R Mills, Ozenberger, Brad A, Ellrott, Kyle, . . . Network, Cancer Genome Atlas Research. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics, 45*(10), 1113.

Wesley-Smith, Ian, & West, Jevin D. (2016). *Babel: A Platform for Facilitating Research in Scholarly Article Discovery*. Paper presented at the Proceedings of the 25th International Conference Companion on World Wide Web, Montréal, Québec, Canada. https://doi.org/10.1145/2872518.2890517