ModEx: A text mining system for extracting mode of regulation of Transcription Factor-gene regulatory interaction

Saman Farahmand[1,2], Todd Riley[1,2] and Kourosh Zarringhalam[1,3*]

[1] Computational Sciences PhD program, University of Massachusetts Boston, Boston, USA

[2] Department of Biology, University of Massachusetts Boston, Boston, USA;

[3] Department of Mathematics, University of Massachusetts Boston, Boston, USA

* Correspondence: kourosh.zarringhalam@umb.edu, Department of Mathematics, University of Massachusetts Boston, 100 Morrissey Boulevard, 02125 Boston, USA

Abstract— Transcription factors (TFs) are proteins that are fundamental to transcription and regulation of gene expression. Each TF may regulate multiple genes and each gene may be regulated by multiple TFs. TFs can act as either activator or repressor of gene expression. This complex network of interactions between TFs and genes underlies many developmental and biological processes and is implicated in several human diseases such as cancer. Hence deciphering the network of TF-gene interactions with information on mode of regulation (activation vs. repression) is an important step toward understanding the regulatory pathways that underlie complex traits. There are many experimental, computational, and manually curated databases of TF-gene interactions. In particular, high-throughput ChIP-seq datasets provide a large-scale map or transcriptional regulatory interactions. However, these interactions are not annotated with information on context and mode of regulation. Such information is crucial to gain a global picture of gene regulatory mechanisms and can aid in developing machine learning models for applications such as biomarker discovery, prediction of response to therapy, and precision medicine. In this work, we introduce a text-mining system to annotate ChIP-seq derived interaction with such meta data through mining PubMed articles. We evaluate the performance of our system using the gold standard small scale manually curated TRUSST database. Our results show that the method is able to accurately extract mode of regulation with

F-score 0.77 on TRRUST curated interaction and F-score 0.96 on intersection of TRUSST and ChIP-network. We provide a HTTP REST API for our code to facilitate usage.

Availability: Source code and datasets are available for download on GitHub:

https://github.com/samanfrm/modex

HTTP REST API: https://watson.math.umb.edu/modex/

Index Terms— Text mining, information extraction, name entity recognition, biological NLP, biomedical literature, gene regulatory network, mode of regulation.

1. INTRODUCTION

Gene regulatory networks are essential in many cellular processes, including metabolism, signal transduction, development, and cell fate [1]. At the transcriptional level, regulations of genes are orchestrated by concerted action between Transcription Factors (TFs), histone modifies, and distal *cis*-regulatory elements to finely tune and modulate expression of genes. Sequence-specific Transcription Factors play a key role in regulating gene transcription at the transcriptional level. They bind specific DNA motifs to regulate promoter activity and either enhance (activate) or repress (inhibit) expression of the genes. Deciphering transcriptional regulatory networks is crucial for understanding cellular mechanisms and response at a molecular level and can shed light on molecular basis of complex human diseases [2–5]. Moreover, knowledge on interactions between genes and biomolecules is an essential building block in several pathway inference and gene enrichment analysis methods that aim to annotate an altered set of transcripts with biological function [6,7]. There are several sources of publicly available biomolecular interactions, including, signaling pathways, metabolic pathways, and protein-protein interactions [8–10]. Databases of transcriptional regulatory network include JASPAR [11], the Open Regulatory Annotation database (ORegAnno) [12] , Swissregulon [13], the Transcriptional Regulatory Element Database (TRED) [14], the Transcription Regulatory Regions Database (TRRD) [15], TFactS [16], TRRUST [17], and the Human Transcriptional Regulation Interactions database (HTRIdb) [17]. These databases include biologically validated and computationally inferred gene regulatory interactions and have been assembled with a variety of approaches, including reverse engineering approaches based on high-throughput gene expression experiments [18–20], text mining approaches [21], and manual curation [22]. These databases are a valuable source of gene regulatory information, however, there are several constraints that limit their usability. For example, databases of computationally predicted and expression-driven interactions are

typically very noisy. Importantly, the majority of the databases do not report the direction of regulation (up or down) - which is crucial to understanding the functional behavior of the cell.

A high-throughput experimental approach for identifying regulatory interaction is chromatin immunoprecipitation followed by sequencing (ChIP-seq). In ChIP-seq methodologies, antibodies that recognizes a specific TF are used to pull down attached DNA for sequencing. The ENCODE (Encyclopedia of DNA Elements) consortium [23] has produced vast amount of publicly available high-throughput ChIP-seq experiments that are processed and deposited into databases such as GTRD [24] and ChIP-Atlas [25] (>35,000 experiments). These databases can be utilized to construct a high coverage transcriptional regulatory network. Although these interactions are experimentally derived, they are still very noisy as the experiments are performed under different conditions and in different cell lines. Moreover, ChIP-seq does not provide direct information on mode of regulation.

The most reliable source of regulatory information is obtained by manual curation of peer-reviewed biomedical literature by domain experts and can be considered as the gold standard. Commercial vendors such as Ingenuity (www.ingenuity.com) offer pathway inference analysis algorithms on such manually curated networks of regulatory interactions. There are also public sources of curated causal gene regulatory interactions, such as TRRD, TRED, TFactS, and TRRUST. However, these databases are very small in their scope, covering only a fraction of TF-gene interactions. Overall, manual curation of biomedical literature is very time consuming, requires extensive resources, and does not scale to the pace at which biomedical literature is expanding [26]. For this reason, biomedical text mining has been extensively used to automate the process of biomolecular relation extraction from the literature. As literature lacks standardized representation of text, automatic routines for information extraction from textual context is very challenging [27].

There is a vast amount of literature on text-mining for various application [27]. Essential text mining steps for biomedical relation extraction can be divided into 3 steps: (1) information retrieval (IR), (2) entity recognition (ER), and (3) information extraction (IE). Together, they can be utilized to identify and extract specific biological knowledge from literature [28,29].

IR tools retrieve relevant text information from articles, abstracts, paragraphs, and sentences corresponding to subject of interest. A popular IR approach for biomedical application is the use of Boolean model logic (AND/OR) for extracting relevant information containing specific biological terms [27]. Prominent IR tools that use the Boolean logic model are iHOP [30] and PubMed. PubMed utilizes human-indexed MeSH terms to reduce the search space and retrieve relevant abstracts containing user specified keywords. iHOP builds on PubMed and is able to detect co-occurrence of terms. A limitation of iHOP is that the terms must occur in the same sentence.

After the IR step, ER must be used to identify relation between biological entities. This is a challenging step as entity names are not unique. Therefore, ER tools must take textual context into consideration to accurately detect entities. For example, gene names may have different variations in ortho-graphical structure (e.g. ABL1, Abl1, Abl-1) or multiple synonyms (e.g. ABL1, ABL, CHDSKM, Abelson tyrosine-protein kinase 1). ER methods, typically divide the task into two steps, (1) identify the entities and their location in the context, and (2) assign unique identifiers to the entities [27]. Fortunately, multiple terminological databases, such as Gene Ontology [31], UMBLS [32], BioLexicon [32], and Biothesaurus [32] provide information on biological entities and name variations and can be used to detect biological entities such as genes or proteins [33–35].

Lastly, Relation Extraction (RE) is an IR tasks for extracting pre-defined facts relating to an entity or entities in the text [36]. In biomedical domain, multiple RE methods have been developed to extract information relating to genes [17], such as Mutation-Disease associations, protein-protein interaction [37,38], pathway curation [39], gene methylation and cancer relation [40], biomolecular events [41], metabolic reactions [42] and gene-gene interactions [43]. For gene regulatory networks, which is the focus of this paper, the RE system must detect and extract a *causal* relation between a protein and a gene (e.g., A regulated B). This task is very complex, even for human experts [44]. To illustrate, consider the causal relation "aatf upregulates c-myc" that should be deduced from the following sentence: "down-regulation of c-myc gene was accompanied by decreased expressions of c-myc effector genes coding for htert, bcl-2, and aatf" [45]. Extracting a positive regulatory interaction between aatf and c-myc is quite challenging using simple RE methods. For example, the RE method, may naively annotated the interaction as negative because of the keyword "decreased". However, by taking "down-regulation" into account, the RE method would able to correctly extract a positive regulation from this sentence.

Construction of a causal transcriptional regulatory network by traditional means of text mining is hampered by these challenges and as a result, fully automated text-mining based models are limited in their scope and accuracy [27]. Combining experimentally-derived regulatory interactions from high-throughput sources with text-mining approaches can bridge the gap between the two approaches and address their shortcomings.

In this work, we present a hybrid model ModEx, to mine the biomedical literature to extract and annotate causal transcriptional regulatory interactions derived from high-throughput ChIP-seq datasets. Specifically, we applied our text-mining method to extract experimental TF-gene relations reported in ChIP-Atlas (assembled from all publicly available ChIP-seq experiments) from biomedical literature and annotated the retrieved interaction with meta-data, such as full supporting sentences, PubMed ID, and importantly mode of regulation, which is missing from ChIP-seq data. It is important to note that our approach bypasses

several of the challenges of fully automated text-mining methods, including query translation for a particular interaction, relevant citation retrieval, entities recognition and regulatory annotation. We evaluated the performance of our model using the TRRUST network [22], which contains 9,396 manually curated regulatory interactions. Our model was able to achieve an F-score 0.77 in retrieving and annotating the TRUSST network. When applied to TRRUST reported interaction that are also present in ChIP-seq data, the method achieved an F-score of 0.96.

2. MATERIALS AND METHODS

We begin by a brief overview of our text mining approach for extracting and annotating ChIP-seq derived TF-gene interactions with meta-data. We acquired all citations from PubMed abstracts by submitting quarries to the database with appropriate Boolean logic regarding entities and their synonyms. State-of-the-art external ER systems such as PubTator [46] and beCAS [47] along with our ER system were utilized to obtain a list of biological entities in the abstract. We then used the Stanford dependency parser [48] to extract dependencies on different sentences and merge the parse trees into a parse graph. The major advantage of this parse graph is its potential to identify long-range dependency relations across sentence boundaries. Candidate relations were created by extracting subtrees connecting pairs of entities from the dependency graph. Finally, we extracted the mode of regulation based on two sets of manually-annotated positive and negative causal categories (consisting > 100 verbs and their inflections). In the subsequent sections we describe the details of our text-mining system.

2.1 Data sets

PubMed database was used to query the entities relating to interaction in ChIP-Atlas. PubMed provides more than 25 million biomedical and life sciences journal articles. TRRUST regulatory network [49] was utilized as gold standard to evaluate the performance of ModEx. TRRUST is a manually-curated database of human transcriptional regulatory network with partial information on mode of regulation. It contains 9,396 regulatory interactions of 800 human transcription factors, 5,066 of which are annotated with information on mode of regulation (3,148 repression and 1,918 activation). We also obtained TF-gene interaction data from ChIP-seq experiments, deposited on the ChIP-Atlas database [25]. This database contains all publicly available high-throughput ChIP-seq experiments. We assembled regulatory networks from these interactions using various cutoff criteria for ChIP-seq peak signal score and distance to the TSS. The least stringent criterion results in a network with 4 million interactions between 758 TFs and 18,874 target genes. There is no reported mode of regulation in this database.

2.2 Extraction of relevant citations

For each regulatory interaction in our assembled ChIP-derived network, we developed an IR system to retrieve the information from the literature. Figure 1 illustrates the overall workflow of our IR component to fetch relevant citations associated with the regulatory interaction. We built a query based on the entities participated in the interactions to retrieve abstracts from PubMed database.

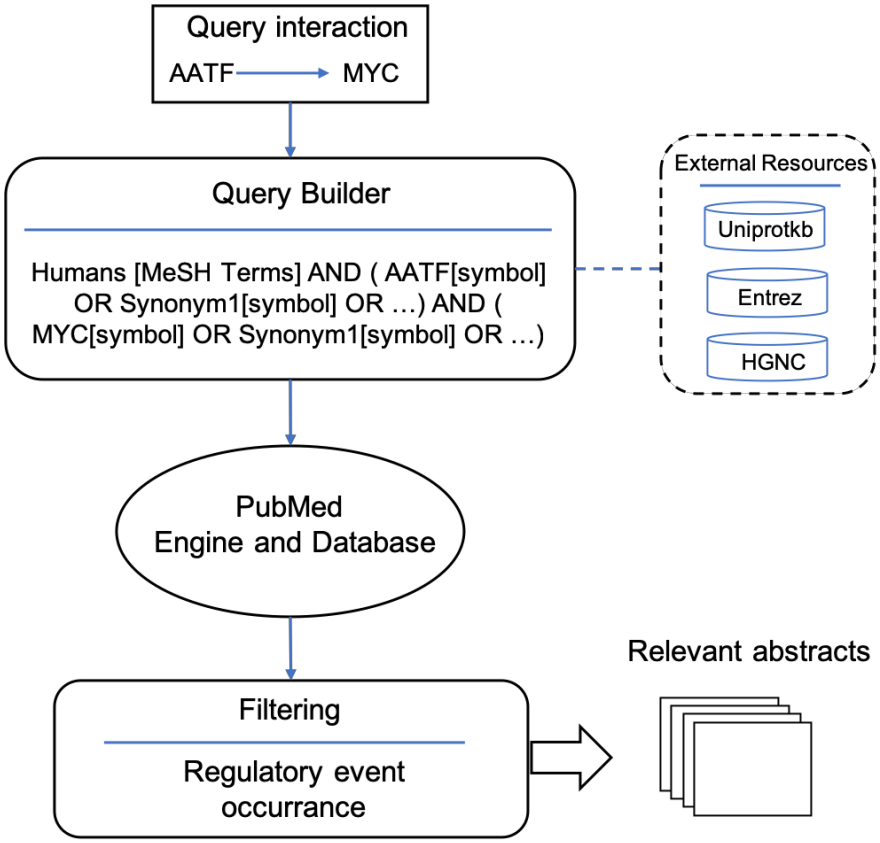


Fig. 1. The Information Retrieval framework. The steps are as follows: first, a Boolean query is built according to the associated entities in the regulatory interaction. It uses several external databases to complement the query with more synonyms and aliases. Then, the query is submitted to the PubMed databased and abstracts are retrieved for processing. Abstracts with no regulatory events are excluded for further analysis.

Each query was supplemented with synonyms acquired from several external resources, including HGNC, Entrez, and UniprotKB to fetch more relevant abstracts. All related citations were acquired by submitting a query with appropriate Boolean logic (AND/OR) on entities and their synonyms. A MeSH descriptor term

173 (e.g. Humans) was also incorporated in the query to reduce the search space. For examples the query for
174 AATF and MYC regulatory interaction is, "humans[mesh terms] AND (AATF[sym] OR BFR2[sym] OR
175 CHE-1[sym] OR CHE1[sym] OR DED[sym]) AND (MYC[sym] OR MRTL[sym] OR MYCC[sym] OR
176 BHLHE39[sym] OR C-MYC[sym])".

177 The queries were submitted via the PubMed engine, a search engine that provides access to the MEDLINE
178 database of references and abstracts on life sciences and biomedical articles. In our implementation we
179 utilized Biopython [50] to run the queries through PubMed engine. We applied a filter on retrieved abstracts
180 and retained only those containing expert-generated "regulatory events" as presented in Table 1. Each
181 category contains more than 50 verbs and their inflections. For example, the ATTF-MYC query outlined
182 above, resulted in 4 relevant abstracts (PMIDs: 20549547, 17006618, 17006618, 20924650).

183

184 **TABLE 1.** Regulatory events categories

Category	NO. Events	Examples
Positive	500	increase, induce, activate, enhance, up-regulate, …
Negative	511	reduce, decrease, suppress, block, down-regulate, decrease, …

185

186 **2.3 Gene and regulatory event recognition**

187 The next step in the pipeline is to identify biological entities within the abstracts. Figure 2 shows the NER
188 component of our system. Two external state-of-the-art NER systems were utilized to annotate the retrieved
189 abstracts with an accurate and complete list of biological entities. The first system is PubTator [51], a web-
190 based system for assisting biocuration. PubTator utilizes a HTTP REST interface, equipped with multiple
191 state-of-the-art text mining algorithms to run queries. Using this system, we queried the retrieved PMIDs
192 and obtained entity annotations in a JSON encoded text. Additionally, we utilized BeCAS [51] (another
193 online NER tool) to improve the coverage of the entities. BeCAS, like PubTator, provides a RESTful API
194 for biomedical name identification. It can run queries directly on provided text or PMIDs and returns
195 associated annotations as an XML document.

196

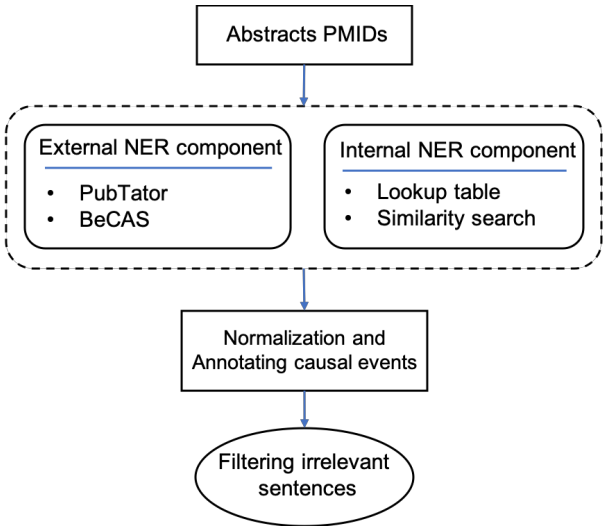


Fig. 2. The Gene and regulatory event recognition workflow. Each PubMed ID retrieved by IR component are submitted to the external NER tools (PubTator and BeCAS) for annotating genes in the abstracts. It follows complementary annotations using our internal NER component including a lookup table for covering acronyms, and a similarity search to identify lexical variations for gene names.

To further enhance the NER system, we implemented and added an additional NER component as follows. Abstracts were normalized to uppercase format and searched for gene acronyms using a manually-curated lookup table [52]. This table includes long term / short term pair association to recognize entities, which were missed by the external NER tools. For instance, AR is a short term for "Androgen Receptor" and was only detected as an entity (transcription factor) using this lookup table. Furthermore, we utilized a name similarity metric to identify strings with lexical variations such as whitespace and punctuations. For instance, "IL-12" and "IL12" are two lexical variations of "Interleukin 12". The former version was not identified by the External NER systems. In our implementation, we set the entity detection threshold based on Jaro similarity [53] of 0.9 or larger between the query entity and the string in the abstract.

Next, we normalized the annotated word or a group of words corresponding to a gene to their HGNC symbol for simplification of downstream analysis. Regulatory events were also annotated using our expert-generated categories (Table 1). Figure 3 illustrates the normalization of gene names and annotation of regulatory events. Sentences that contained no regulatory event were excluded from further analysis.

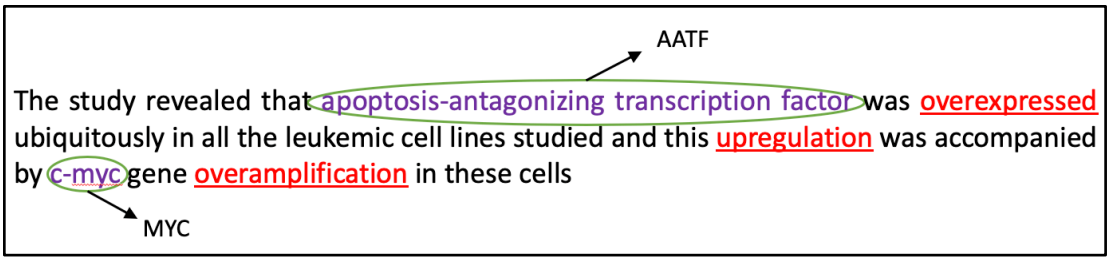


Fig. 3. An example of gene entity normalization and regulatory events annotation. All of the words or group of words associated to target entities (purple color) are normalized to their HGNC symbol for simplification. Causal regulatory events also are annotated according to their categories, and sentences with no regulatory event are excluded for further consideration.

2.4. Extracting mode of regulation

For each causal interaction, its associated annotated sentences within relevant abstracts were submitted to the Stanford dependency parser [48] and a dependency parse tree was generated. Dependency trees extracted from different sentences were merged into a single large graph. The merging process is straightforward; Each dependency relation includes one head word/node and one dependent word/node. Nodes from different dependency relations representing the same word were. PMID was recorded for each edge in the parse tree to indicate its source. Each edge in the parse tree was assigned a weight based on the number of occurrences of dependency relations. The rational for using this weighted parse tree is that it can be used to identify long-range dependency relations across sentence boundaries that would otherwise be missed. Figure 4 shows the relation extraction workflow of our method. Absolute frequency of a dependency relation obtained from the merging step can somewhat reflect the semantic relation of the head word and the dependent word.

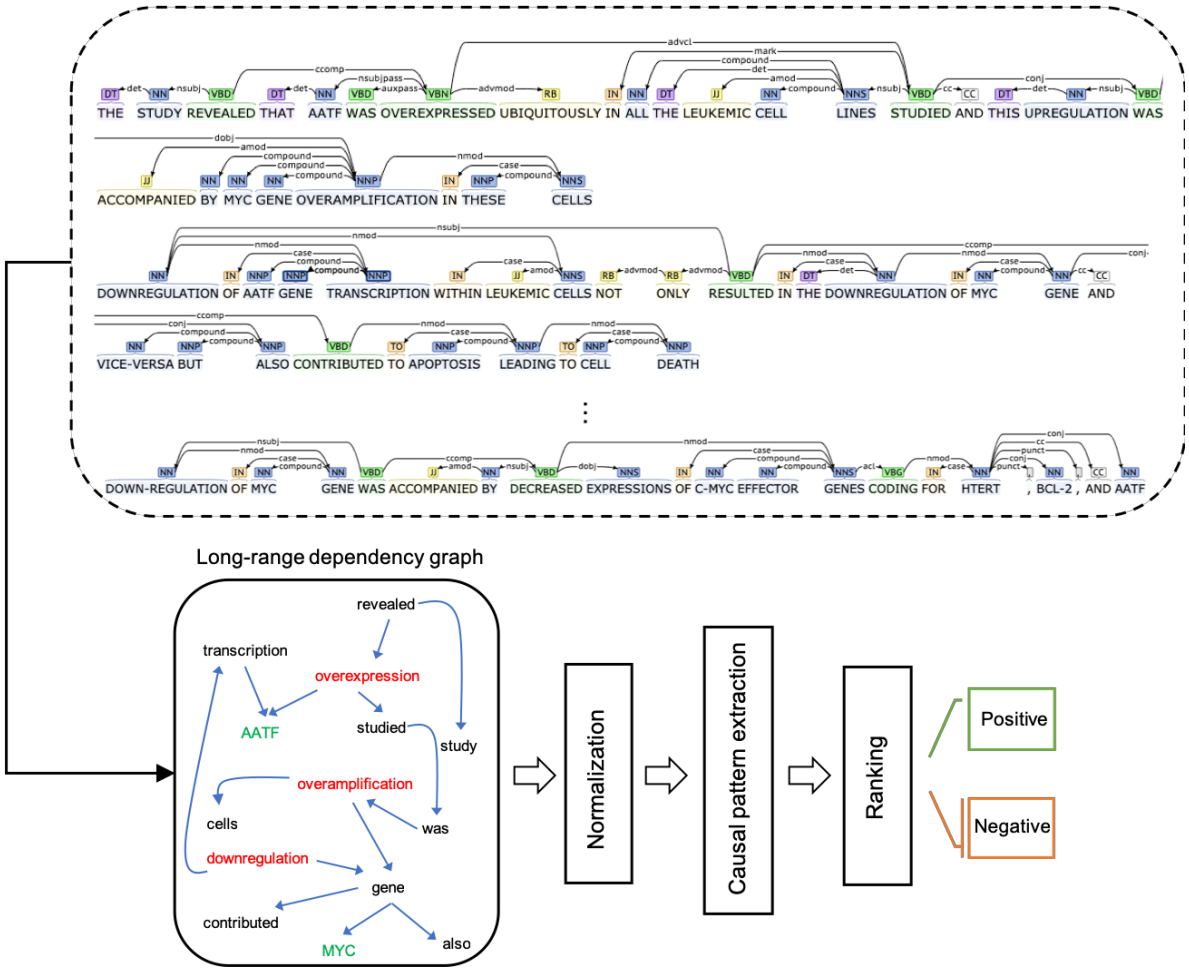


Fig. 4. Regulatory sign extraction workflow. A long-range dependency graph is constructed by merging all of dependency trees corresponding to the evidence sentences. The weights of the graph are the number of occurrences of dependency relations. Candidate regulatory signs are identified using common subtrees with at least one regulatory event in the graph. Finally, a sign of regulation is assigned to the query interaction through the ranking task.

Our system, ModEx, creates candidate relations by extracting subtrees with common ancestors connecting the pair of query genes as leaves. These subtrees must contain at least one causal event describing the candidate relation between the given pair of genes. Subtrees were extracted by applying a Depth First Search along with a boolean visited array to avoid possible loops. Nodes with two paths to the entities were considered as a root of the subtree. Next, we utilized a rule based approach to describe relations using three commonly used language constructs [36]. The first rule is effector-relation-effectee (e.g. A activates B). The second rule is relation-of-effectee-by-effector (e.g. Activation of A by B). These rules were applied to both paths from root to query entities to identify their regulatory dependency. Figure 5a illustrates the regulatory relation extraction using these rules. Some sentences in the literature have complex structures,

which cannot be captured by these language constructs. To address this, we also incorporated a negation rule to increase the performance of the RE system. For example, consider the following sentence: "LMP1 suppresses the transcriptional repressor ATF3, possibly leading to the TGFβ-induced ID1 upregulation" [54]. In the first pass the system assigns a positive mode to the interaction between ATF3 and ID1. However, there is a negative interaction between the TF and target gene. The negation rule considers the negative event "suppresses" related to ATF3 and switch the positive mode to negative. Figure 5b shows a subtree reflecting the negation rule.

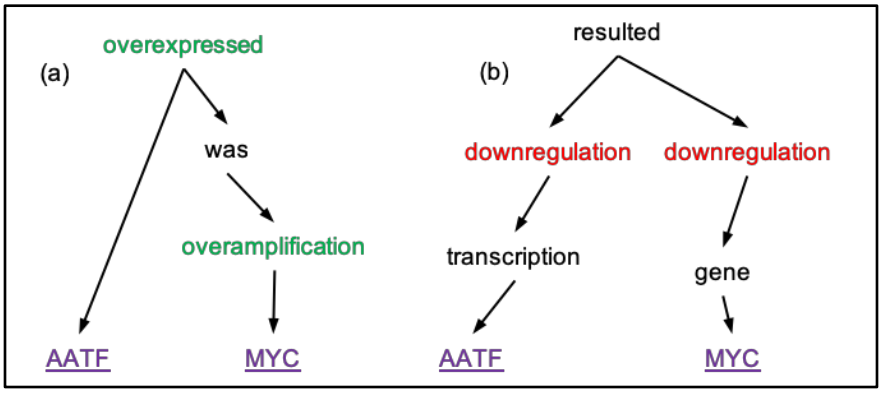


Fig. 5. Examples to illustrate the rules for finding the regulatory sign. panel (a) shows an example for simple rule (effector-relation-effectee) in which the RE system can assign a positive sign to this candidate pattern. In panel (b), we can see the impact of the negation rule to extract accurate sign to this pattern. Two paths from root to query entities contain negative regulatory events which carries an activation/positive sign for the pattern.

We then ranked each subtree based on the sign of regulatory interaction between the query genes. The weights of the graph encode repetition of regulatory relations across sentences and abstracts. we considered the weights when there were more than one regulatory event associated with the target gene. In this case, an event with higher weight was selected for ranking the subtree. We also considered distance of events to the target gene when the weights in the subtree were equal. The closest event to the target entity will take the highest priority for determining the interaction sign. Finally, we investigated signs in every candidate subtree and assigned a total sign of regulation to the interaction using a voting scheme.

2.5. ModEx HTTP Interface

We implemented an HTTP REST server for users to programmatically annotate gene regulatory networks using ModEx. Clients should make HTTP requests to the server with a particular format, specifying the query entities and optional MeSH term to annotate. The query has to be requested in the following format: *TFEntrezID_TargetEntrezID_MeSHterm[optional].* For instance, a query to the server for AATF-MYC should be formatted as "*/signex/26574_4609_humans".* The server returns extracted annotation along with associated citations and sentences from PubMed database if any evidence exist. The server can be accessed at: https://watson.math.umb.edu/modex/

3 RESULTS

3.1 Classification Performance

We evaluated the performance of our method using the TRRUST database, a manually-curated network or regulatory interaction with partial information on mode of regulation. TRRUST is a high-quality database and can be considered as gold standard for our benchmark. We applied our method to 5,066 regulatory interactions in TRRUST for which information on mode of regulation was available. Table 2 shows the summary statistic of the results.

TABLE 2. Summary statistics of performing ModEx on TRRUST

System	Without evidence	With evidence	Detected with sign	
			4225	
modEx	182	4884	Positive	Negative
			2659	1566

Our method did not detect any PubMed abstracts for 182 of interactions. ModEx detected 4,225 signs corresponding to 4,225 regulatory interactions including 2,659 positive and 1,566 negative interactions. We compare the identified signs by ModEx with reported signs in the TRRUST database. Our system correctly extracted 2,216 positive and 1,024 negative signs with overall accuracy of 0.76. Figure 6 shows the

classification result of ModEx on the TRRUST database using various metrics (Precision, Recall and F-score).

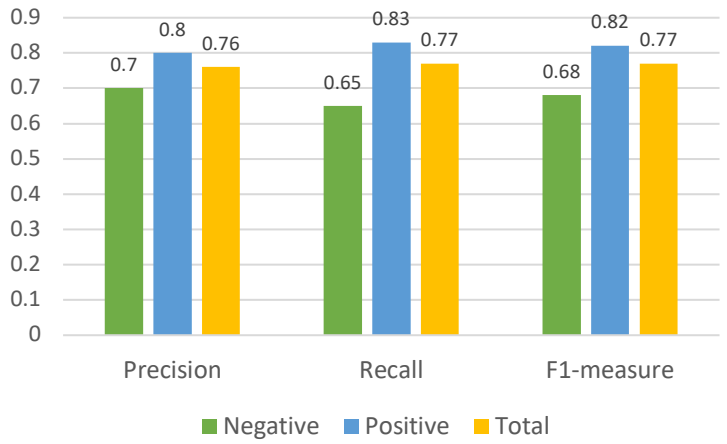


Fig. 6. Classification results of ModEx on TRRUST

We also compared the citations of the 4,884 retrieved by our system with the citations reported in the TRRUST database. All citations match, with accuracy of 1.0. This validation using the gold standard demonstrates the ability of our system to correctly identify relevant citations, extract causal interactions between TF and gene, and detect mode of regulation. We used our system to annotate the remining interactions in the TRUSST database for which no mode of regulation is reported.

3.2 ChIP-Atlas Analysis

We next sought to extract and annotate ChIP-seq derived TF-gene causal regulatory interactions from literature using our system. Such meta-data and evidence from literature can increase the confidence in the TF-gene interactions identified by ChIP-seq experiments and further shed light on the mechanism of interaction. Information on mode of regulation in particular can be helpful to enhance the accuracy of enrichment algorithms for regulatory pathway inference [55].

We applied ModEx to ChIP-seq interactions, with moderately stringency criteria, i.e., binding distance within 1k of the TSS and ChIP peak score > 950, resulting in 43,444 interactions. The system was able to detect and annotate 1,592 of interactions in PubMed database. Table 3 outlines the results.

316 **TABLE 3.** Summary statistics of performing ModEx on ChIP-Atlas

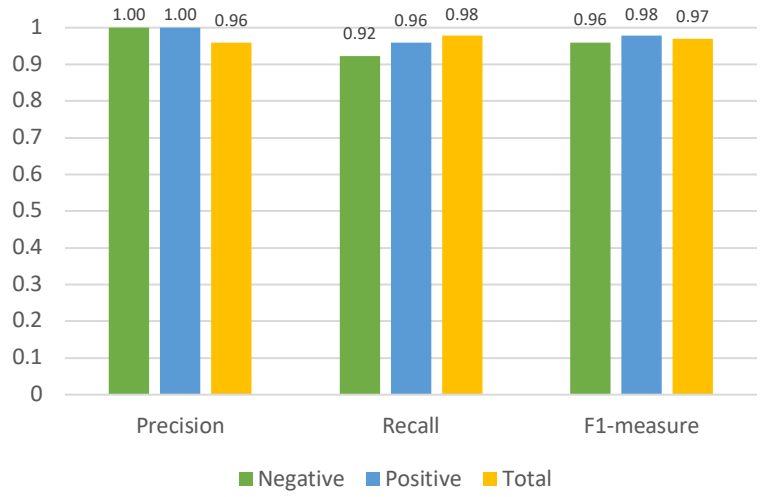
System	Overall	With evidence	Detected with sign	
			1,592	
ModEx	43,444	5,133	Positive	Negative
			1,421	171

317

318 Some of the retrieved annotated ChIP-seq interactions also appear in the TRRUST database (69 total). We
319 compared the identified mode of regulations of ChIP-seq interactions with the reported signs in the
320 TRRUST database. Figure 7 summarizes the classification results. As can be seen the agreement is very
321 high, indicating that our method can reliably identify and annotate ChIP interaction when they are reported
322 in literature. Additionally, we compared our acquired evidence (PMIDs) by ModEx with citations reported
323 in TRRUST. Our IR module was able to fetch the relevant evidence from PubMed database with accuracy
324 0.88.

325

326



327

328

329 **Fig. 7.** Classification results of ModEx on intersctio on TRRUST and ChIP-Atlas

330 ### 3.3 Directional enrichment analysis

331 To demonstrate the utility of our annotated network, we used our network in conjunction with a directional
332 enrichment analysis algorithm [55] to identify drivers of differential expressed genes. We utilized
333 quaternaryProd, a gene set enrichment algorithm that can take advantage of direction of regulation on causal
334 biological interaction graphs to identify regulators of differential gene expression. The algorithm can take
335 a signed transcriptional regulatory network, such as TRRUST or our annotated ChIP-network along with a
336 differential expression profile as input and outputs a set of candidate active regulators. The ability of the
337 algorithm to identify regulators of differential gene expression relies heavily on the quality and the coverage
338 of the regulatory network on which the queries are performed. To test the utility of our network, we used
339 this algorithm along with differential expression profiles from controlled over-expression experiments used
340 in the original study. The over-expression experiments consist of differential gene expression profile from
341 a controlled in vitro E2F3 over expression [56] and c-Myc [56]. We inputted three networks into the
342 algorithm (1) the original TRUSST network, (2) annotated TRUSST network, and (3) annotated TRRUST
343 augmented with annotated ChIP-Atlas. By annotated TRRUST, we refer to the TRRUST network where
344 interaction with no reported mode of regulation were annotated using our system. Differential gene
345 expression analysis of these data sets resulted in 272, and 220 differentially expressed genes
346 respectively. Table 4 outlines the top 5 regulators predicted by the algorithm on E2F3 experiment sorted
347 by the FDR corrected *p-values* of the scoring scheme. For the E2F3 experiment, E2F1 is returned as the top
348 hypothesis regulator by the algorithm incorporating our annotated networks. E2F1 and E2F3 are close
349 family members and have a very similar role as transcription factors that function to control the cell cycle
350 and are similarly implicated in cancer [57]. It is interesting to note that original TRRUST database does not
351 include enough information for algorithm to recover E2F1, however the signal strengthens when TRUSST
352 is annotated with our system and a much more significant p-value is obtained when TRRUST is augmented
353 with annotated ChIP-Atlas. This shows that annotating ChIP-seq data provides significant additional power
354 to identify upstream regulators in conjunction with freely available causal networks.

355

356 **TABLE 4.** Directional enrichment analysis results on E2F1 expression signatures

TRRUST			**Annotated TRRUST**			**Annotated TRRUST with ChIP-Atlas**		
Name	**Regulation**	**Adj. P-Val**	**Name**	**Regulation**	**Adj. P-Val**	**Name**	**Regulation**	**Adj. P-Val**
REL	Down	1.5e-3	**E2F1**	**Up**	**1.2e-5**	**E2F1**	**Up**	**1.2e-7**
PROX1	Up	2.9e-3	PROX1	Up	2.1e-3	PROX1	Up	2.1e-3
SUGP1	Down	3.2e-3	SUGP1	Down	2.4e-3	SUGP1	Down	2.4e-3

NFIL3	Up	6.3e-3	RELA	Down	3.5e-3	RELA	Down	3.5e-3
TFDP1	Up	7e-3	TFDP1	Up	3.9e-3	TFDP1	Up	3.9e-3

Application of the method to c-Myc differential expression profile shows the same pattern. The annotated TRRUST with ChIP-Atlas recovered MAX as one of the top 20 regulators with low p-value compared to TRRUST. It has been demonstrated that oncogenic activity of c-Myc requires dimerization with MAX [58].

TABLE 5. Directional enrichment analysis results on c-Myc expression signatures

TRRUST			**Annotated TRRUST**			**Annotated TRRUST with ChIP-Atlas**		
Name	**Regulation**	**Adj. P-Val**	**Name**	**Regulation**	**Adj. P-Val**	**Name**	**Regulation**	**Adj. P-Val**
MYBL2	Up	3.6e-3	MAFA	Down	1.3e-2	USF2	Up	8.8e-3
MXI1	Down	4.0e-3	MKL1	Down	1.3e-2	MAFA	Down	1.3e-2
AATF	Up	4.0e-3	GLI3	Down	1.7e-2	MKL1	Down	1.3e-2
ENO1	Down	4.0e-3	KAT2B	Up	1.9e-2	GLI3	Down	1.7e-2
NR1D1	Up	6.4e-3	SOX6	Down	1.9e-2	KAT2B	Up	1.9e-2
TLE3	Up	6.4e-3	HDAC1	Down	2.5e-2	SOX6	Down	1.9e-2
TOP2B	Down	6.4e-3	MYBL2	Down	2.6e-2	HDAC1	Down	2.5e-2
L3MBTL1	Up	6.4e-3	HDAC7	Up	3.0e-2	MYBL2	Down	2.6e-2
MAFA	Down	6.4e-3	ILF3	Down	3.0e-2	HDAC7	Up	3.0e-3
TLX1	Up	7.9e-3	ELK1	Up	3.6e-2	ILF3	Down	3.0e-2
HLF	Up	1.0e-2	GATA6	Up	3.9e-2	ELK1	Up	3.6e-2
DLX5	Up	1.1e-2	PPARG	Down	4.3e-2	GATA6	Up	3.9e-2
HOXA1	Up	1.1e-2	SATB1	Up	4.4e-2	PLAG1	Up	4.0e-2
MKL1	Down	1.2e-2	ARNT	Up	4.8e-2	CREB1	Up	4.1e-2
IFI16	Down	1.4e-2	RXRA	Up	5.3e-2	PPARG	Down	4.3e-2
PRDM1	Down	1.5e-2	ZEB1	Down	5.5e-2	SATB1	Up	4.4e-2
CEBPE	Down	1.6e-2	E2F4	Down	5.5e-2	ZNF143	Up	4.6e-2
HDAC1	Down	1.6e-2	PPARD	Down	5.6e-2	ARNT	UP	4.8e-2
GLI3	Down	1.6e-2	XBP1	Up	6.5e-2	**MAX**	**Up**	**5.3e-2**
STAT4	Up	1.8e-2	DDIT3	Down	6.5e-2	RXRA	Up	5.4e-2

Conclusion

In this work we presented a fully automated text-mining system to extract and annotate causal regulatory interaction between transcription factors and genes from the biomedical literature. As a starting point, our method uses putative TF-gene interactions derived from high-throughput ChIP-seq or other experiments and seeks to collect evidence and meta-data in the biomedical literature to support the interaction. It should be noted that annotating a priori known interactions differs significantly in scope and complexity from general text-mining approaches for biomedical relation extraction. The later attempts to extract the causal relation from biomedical text directly, without prior knowledge of the entities and the interaction, whereas in our method the relation is know from biological experiments and curated databases a priori, thereby reducing the complexity significantly. This approach bridges the gap between data-driven methods and text-mining methods for constructing causal transcriptional gene regulatory networks and overcomes some of the drawbacks of either approach. With the rapid increase in high-throughput experiments and biomedical literature, hybrid method such as the one proposed can make a significant impact in biological knowledge retrieval.

We used a gold-standard manually curated dataset and demonstrated that our approach can reliably identify the relevant literature and extract the correct interaction and meta-data. We applied our method to high-throughput ChIP-seq data and provided literature support for ~1,500 interactions. Our annotated ChIP-derived transcriptional regulatory interaction can be used in conjunction with directional enrichment methods that aim to identify regulators of differential gene expression. Moreover, we use our system to annotate the interactions in the TRRUST database for which more of regulation is not reported. Our system can also be used as a tool to mine the literature for investigate interactions in newly performed ChIP-seq experiments, where researchers are interested to investigate a specific interaction between a protein and a gene. To facilitate usage, we implemented an HTTP REST server for users to programmatically annotate gene regulatory networks using ModEx available to download at: https://watson.math.umb.edu/modex/. The annotated ChIP-network as well as annotated TRRUST can be obtained from: https://doi.org/10.6084/m9.figshare.8251502.v1

ACKNOWLEDGMENT

The authors are grateful to Dr. Nurit Haspel for helpful discussions and comments.

REFERENCES

[1] H. de Jong, Modeling and Simulation of Genetic Regulatory Systems: A Literature Review, J. Comput. Biol. 9 (2002) 67–103. doi:10.1089/10665270252833208.

[2] G. Karlebach, R. Shamir, Modelling and analysis of gene regulatory networks, Nat. Rev. Mol. Cell Biol. 9 (2008) 770–780. doi:10.1038/nrm2503.

[3] S. Farahmand, S. Goliaei, N. Ansari-Pour, Z. Razaghi-Moghadam, GTA: a game theoretic approach to identifying cancer subnetwork markers, Mol. Biosyst. 12 (2016) 818–825. doi:10.1039/C5MB00684H.

[4] S. Farahmand, M.H. Foroughmand-Araabi, S. Goliaei, Z. Razaghi-Moghadam, CytoGTA: A cytoscape plugin for identifying discriminative subnetwork markers using a game theoretic approach, PLoS One. 12 (2017) e0185016. doi:10.1371/journal.pone.0185016.

[5] Z. Razaghi-Moghadam, A. Namipashaki, S. Farahmand, N. Ansari-Pour, Systems genetics of nonsyndromic orofacial clefting provides insights into its complex aetiology, Eur. J. Hum. Genet. 27 (2019) 226–234. doi:10.1038/s41431-018-0263-7.

[6] M. Miwa, T. Ohta, R. Rak, A. Rowley, D.B. Kell, S. Pyysalo, S. Ananiadou, A method for integrating and ranking the evidence for biochemical pathways by mining reactions from text, Bioinformatics. 29 (2013) i44–i52. doi:10.1093/bioinformatics/btt227.

[7] S. Bagewadi, T. Bobić, M. Hofmann-Apitius, J. Fluck, R. Klinger, Detecting miRNA Mentions and Relations in Biomedical Literature, F1000Research. 3 (2015) 205. doi:10.12688/f1000research.4591.3.

[8] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K.P. Tsafou, M. Kuhn, P. Bork, L.J. Jensen, C. von Mering, STRING v10: protein–protein interaction networks, integrated over the tree of life, Nucleic Acids Res. 43 (2015) D447–D452. doi:10.1093/nar/gku1003.

[9] M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Res. 28 (2000) 27–30. doi:10.1093/nar/28.1.27.

[10] E.G. Cerami, B.E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G.D. Bader, C. Sander, Pathway Commons, a web resource for biological pathway data, Nucleic Acids Res. 39 (2011) D685–D690. doi:10.1093/nar/gkq1039.

[11] A. Sandelin, W. Alkema, P. Engström, W.W. Wasserman, B. Lenhard, JASPAR: an open-access database for eukaryotic transcription factor binding profiles, Nucleic Acids Res. 32 (2004) 91D – 94. doi:10.1093/nar/gkh012.

[12] O.L. Griffith, S.B. Montgomery, B. Bernier, B. Chu, K. Kasaian, S. Aerts, S. Mahony, M.C. Sleumer, M. Bilenky, M. Haeussler, M. Griffith, S.M. Gallo, B. Giardine, B. Hooghe, P. Van Loo, E. Blanco, A. Ticoll, S. Lithwick, E. Portales-Casamar, I.J. Donaldson, G. Robertson, C. Wadelius, P. De Bleser, D. Vlieghe, M.S. Halfon, W. Wasserman, R. Hardison, C.M. Bergman, S.J.M. Jones, ORegAnno: an open-access community-driven resource for regulatory annotation, Nucleic Acids Res. 36 (2007) D107–D113. doi:10.1093/nar/gkm967.

[13] M. Pachkov, I. Erb, N. Molina, E. van Nimwegen, SwissRegulon: a database of genome-wide annotations of regulatory sites, Nucleic Acids Res. 35 (2007) D127–D131. doi:10.1093/nar/gkl857.

[14] C. Jiang, Z. Xuan, F. Zhao, M.Q. Zhang, TRED: a transcriptional regulatory element database, new entries and other development, Nucleic Acids Res. 35 (2007) D137–D140. doi:10.1093/nar/gkl1041.

[15] N.A. Kolchanov, O.A. Podkolodnaya, E.A. Ananko, E. V. Ignatieva, I.L. Stepanenko, O. V. Kel-Margoulis, A.E. Kel, T.I. Merkulova, T.N. Goryachkovskaya, T. V. Busygina, F.A. Kolpakov, N.L. Podkolodny, A.N. Naumochkin, I.M. Korostishevskaya, A.G. Romashchenko, G.C.

Overton, Transcription Regulatory Regions Database (TRRD): its status in 2000, Nucleic Acids Res. 28 (2000) 298–301. doi:10.1093/nar/28.1.298.

[16] A. Essaghir, F. Toffalini, L. Knoops, A. Kallin, J. van Helden, J.-B. Demoulin, Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data, Nucleic Acids Res. 38 (2010) e120–e120. doi:10.1093/nar/gkq149.

[17] H. Han, H. Shim, D. Shin, J.E. Shim, Y. Ko, J. Shin, H. Kim, A. Cho, E. Kim, T. Lee, H. Kim, K. Kim, S. Yang, D. Bae, A. Yun, S. Kim, C.Y. Kim, H.J. Cho, B. Kang, S. Shin, I. Lee, TRRUST: a reference database of human transcriptional regulatory interactions, Sci. Rep. 5 (2015) 11432. doi:10.1038/srep11432.

[18] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, N. Friedman, Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, Nat. Genet. 34 (2003) 166–176. doi:10.1038/ng1165.

[19] S. Farahmand, S. Goliaei, Z.R.M. Kashani, S. Farahmand, Identifying Cancer Subnetwork Markers Using Game Theory Method, in: Springer, Singapore, 2019: pp. 105–109. doi:10.1007/978-981-10-4505-9_17.

[20] A. Vajdi, N. Haspel, H. Banaee, A new DP algorithm for comparing gene expression data using geometric similarity, in: 2015 IEEE Int. Conf. Bioinforma. Biomed., IEEE, 2015: pp. 1157–1161. doi:10.1109/BIBM.2015.7359846.

[21] M. Krallinger, A. Valencia, Text-mining and information-retrieval services for molecular biology, Genome Biol. 6 (2005) 224. doi:10.1186/gb-2005-6-7-224.

[22] H. Han, J.-W. Cho, S. Lee, A. Yun, H. Kim, D. Bae, S. Yang, C.Y. Kim, M. Lee, E. Kim, S. Lee, B. Kang, D. Jeong, Y. Kim, H.-N. Jeon, H. Jung, S. Nam, M. Chung, J.-H. Kim, I. Lee, TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions, Nucleic Acids Res. 46 (2018) D380–D386. doi:10.1093/nar/gkx1013.

[23] T.E.P. ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project., Science. 306 (2004) 636–40. doi:10.1126/science.1105136.

[24] I. Yevshin, R. Sharipov, T. Valeev, A. Kel, F. Kolpakov, GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments, Nucleic Acids Res. 45 (2017) D61–D67. doi:10.1093/nar/gkw951.

[25] S. Oki, T. Ohta, G. Shioi, H. Hatanaka, O. Ogasawara, Y. Okuda, H. Kawaji, R. Nakaki, J. Sese, C. Meno, ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data, EMBO Rep. 19 (2018) e46255. doi:10.15252/embr.201846255.

[26] S. Gebel, R.B. Lichtner, B. Frushour, W.K. Schlage, V. Hoang, M. Talikka, A. Hengstermann, C. Mathis, E. Veljkovic, M. Peck, M.C. Peitsch, R. Deehan, J. Hoeng, J.W. Westra, Construction of a computable network model for DNA damage, autophagy, cell death, and senescence., Bioinform. Biol. Insights. 7 (2013) 97–117. doi:10.4137/BBI.S11154.

[27] L.J. Jensen, J. Saric, P. Bork, Literature mining for the biologist: from information retrieval to biological discovery, Nat. Rev. Genet. 7 (2006) 119–129. doi:10.1038/nrg1768.

[28] A.M. Cohen, W.R. Hersh, A survey of current work in biomedical text mining, Brief. Bioinform. 6 (2005) 57–71. doi:10.1093/bib/6.1.57.

[29] P. Zweigenbaum, D. Demner-Fushman, H. Yu, K.B. Cohen, Frontiers of biomedical text mining: current progress, Brief. Bioinform. 8 (2007) 358–375. doi:10.1093/bib/bbm045.

[30] R. Hoffmann, A. Valencia, Implementing the iHOP concept for navigation of biomedical

485 literature, Bioinformatics. 21 (2005) ii252–ii258. doi:10.1093/bioinformatics/bti1142.
486 [31] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K.
487 Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis,
488 J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene Ontology: tool
489 for the unification of biology, Nat. Genet. 2000 251. (2000).
490 [32] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical
491 terminology., Nucleic Acids Res. 32 (2004) D267-70. doi:10.1093/nar/gkh061.
492 [33] A. Yeh, A. Morgan, M. Colosimo, L. Hirschman, BioCreAtIvE Task 1A: gene mention finding
493 evaluation, BMC Bioinformatics. 6 (2005) S2. doi:10.1186/1471-2105-6-S1-S2.
494 [34] Y. Mao, K. Van Auken, D. Li, C.N. Arighi, P. McQuilton, G.T. Hayman, S. Tweedie, M.L.
495 Schaeffer, S.J.F. Lauledеrkind, S.-J. Wang, J. Gobeill, P. Ruch, A.T. Luu, J. -j. Kim, J.-H. Chiang,
496 Y.-D. Chen, C.-J. Yang, H. Liu, D. Zhu, Y. Li, H. Yu, E. Emadzadeh, G. Gonzalez, J.-M. Chen,
497 H.-J. Dai, Z. Lu, Overview of the gene ontology task at BioCreative IV, Database. 2014
498 (2014) bau086–bau086. doi:10.1093/database/bau086.
499 [35] A.A. Morgan, Z. Lu, X. Wang, A.M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman,
500 J. Hakenberg, C. Sun, H. Liu, R. Torres, M. Krauthammer, W.W. Lau, H. Liu, C.-N. Hsu, M.
501 Schuemie, K.B. Cohen, L. Hirschman, Overview of BioCreative II gene normalization.,
502 Genome Biol. 9 Suppl 2 (2008) S3. doi:10.1186/gb-2008-9-s2-s3.
503 [36] K. Fundel, R. Kuffner, R. Zimmer, RelEx--Relation extraction using dependency parse trees,
504 Bioinformatics. 23 (2007) 365–371. doi:10.1093/bioinformatics/btl616.
505 [37] L. Qian, G. Zhou, Tree kernel-based protein–protein interaction extraction from biomedical
506 literature, J. Biomed. Inform. 45 (2012) 535–543. doi:10.1016/J.JBI.2012.02.004.
507 [38] A. Vajdi, K. Zarringhalam, N. Haspel, Patch-DCA: Improved Protein Interface Prediction by
508 utilizing Structural Information and Clustering DCA scores, BioRxiv. (2019) 656074.
509 doi:10.1101/656074.
510 [39] K.E. Ravikumar, K.B. Wagholikar, D. Li, J.-P. Kocher, H. Liu, Text mining facilitates database
511 curation - extraction of mutation-disease associations from Bio-medical literature, BMC
512 Bioinformatics. 16 (2015) 185. doi:10.1186/s12859-015-0609-x.
513 [40] Y.-C. Fang, P.-T. Lai, H.-J. Dai, W.-L. Hsu, MeInfoText 2.0: gene methylation and cancer
514 relation extraction from biomedical literature, BMC Bioinformatics. 12 (2011) 471.
515 doi:10.1186/1471-2105-12-471.
516 [41] M. Gerner, F. Sarafraz, C.M. Bergman, G. Nenadic, BioContext: an integrated text mining
517 system for large-scale extraction and contextualization of biomolecular events,
518 Bioinformatics. 28 (2012) 2154–2161. doi:10.1093/bioinformatics/bts332.
519 [42] J. Czarnecki, I. Nobeli, A.M. Smith, A.J. Shepherd, A text-mining system for extracting
520 metabolic reactions from full-text articles, BMC Bioinformatics. 13 (2012) 172.
521 doi:10.1186/1471-2105-13-172.
522 [43] E.K. Mallory, C. Zhang, C. Ré, R.B. Altman, Large-scale extraction of gene interactions from
523 full-text literature using DeepDive, Bioinformatics. 32 (2015) btv476.
524 doi:10.1093/bioinformatics/btv476.
525 [44] R.J. Mooney, R. Bunescu, Mining knowledge from text using information extraction, ACM
526 SIGKDD Explor. Newsl. 7 (2005) 3–10. doi:10.1145/1089815.1089817.
527 [45] S. Bhatia, D. Kaul, N. Varma, Potential tumor suppressive function of miR-196b in B-cell
528 lineage acute lymphoblastic leukemia, Mol. Cell. Biochem. 340 (2010) 97–106.

529 doi:10.1007/s11010-010-0406-9.
530 [46] C.-H. Wei, H.-Y. Kao, Z. Lu, PubTator: a web-based text mining tool for assisting biocuration,
531 Nucleic Acids Res. 41 (2013) W518–W522. doi:10.1093/nar/gkt441.
532 [47] T. Nunes, D. Campos, S. Matos, J.L. Oliveira, BeCAS: biomedical concept recognition
533 services and visualization, Bioinformatics. 29 (2013) 1915–1916.
534 doi:10.1093/bioinformatics/btt317.
535 [48] D. Klein, C.D. Manning, Fast Exact Inference with a Factored Model for Natural Language
536 Parsing, in: S. Becker, S. Thrun, K. Obermayer (Eds.), Adv. Neural Inf. Process. Syst. 15, MIT
537 Press, 2003: pp. 3–10. http://papers.nips.cc/paper/2325-fast-exact-inference-with-a-
538 factored-model-for-natural-language-parsing.pdf.
539 [49] H. Han, J.-W. Cho, S. Lee, A. Yun, H. Kim, D. Bae, S. Yang, C.Y. Kim, M. Lee, E. Kim, S. Lee, B.
540 Kang, D. Jeong, Y. Kim, H.-N. Jeon, H. Jung, S. Nam, M. Chung, J.-H. Kim, I. Lee, TRRUST v2:
541 an expanded reference database of human and mouse transcriptional regulatory
542 interactions, Nucleic Acids Res. 46 (2018) D380–D386. doi:10.1093/nar/gkx1013.
543 [50] P.J.A. Cock, T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke, I. Friedberg, T.
544 Hamelryck, F. Kauff, B. Wilczynski, M.J.L. de Hoon, Biopython: freely available Python tools
545 for computational molecular biology and bioinformatics, Bioinformatics. 25 (2009) 1422–
546 1423. doi:10.1093/bioinformatics/btp163.
547 [51] C.-H. Wei, H.-Y. Kao, Z. Lu, PubTator: a web-based text mining tool for assisting biocuration,
548 Nucleic Acids Res. 41 (2013) W518–W522. doi:10.1093/nar/gkt441.
549 [52] K.E. Ravikumar, M. Rastegar-Mojarad, H. Liu, BELMiner: adapting a rule-based relation
550 extraction system to extract biological expression language statements from bio-medical
551 literature evidence sentences, Database. 2017 (2017). doi:10.1093/database/baw156.
552 [53] M.A. Jaro, Advances in Record-Linkage Methodology as Applied to Matching the 1985
553 Census of Tampa, Florida, J. Am. Stat. Assoc. 84 (1989) 414. doi:10.2307/2289924.
554 [54] A.K. Lo, C.W. Dawson, K.W. Lo, Y. Yu, L.S. Young, Upregulation of Id1 by Epstein-Barr Virus-
555 encoded LMP1 confers resistance to TGFβ-mediated growth inhibition, Mol. Cancer. 9
556 (2010) 155. doi:10.1186/1476-4598-9-155.
557 [55] C.T. Fakhry, P. Choudhary, A. Gutteridge, B. Sidders, P. Chen, D. Ziemek, K. Zarringhalam,
558 Interpreting transcriptional changes using causal graphs: new methods and their practical
559 utility on public networks, BMC Bioinformatics. 17 (2016) 318. doi:10.1186/s12859-016-
560 1181-8.
561 [56] A.H. Bild, G. Yao, J.T. Chang, Q. Wang, A. Potti, D. Chasse, M.-B. Joshi, D. Harpole, J.M.
562 Lancaster, A. Berchuck, J.A. Olson, J.R. Marks, H.K. Dressman, M. West, J.R. Nevins,
563 Oncogenic pathway signatures in human cancers as a guide to targeted therapies, Nature.
564 439 (2006) 353–357. doi:10.1038/nature04296.
565 [57] H.-Z. Chen, S.-Y. Tsai, G. Leone, Emerging roles of E2Fs in cancer: an exit from cell cycle
566 control, Nat. Rev. Cancer. 9 (2009) 785–797. doi:10.1038/nrc2696.
567 [58] B. Amati, M.W. Brooks, N. Levy, T.D. Littlewood, G.I. Evan, H. Land, Oncogenic activity of
568 the c-Myc protein requires dimerization with Max, Cell. 72 (1993) 233–245.
569 doi:10.1016/0092-8674(93)90663-B.
570