Distant Supervision for Cancer Pathway Extraction from Text

Hoifung Poon*, Kristina Toutanova, Chris Quirk
Microsoft Research, Redmond, WA, USA
**E-mail: hoifung@microsoft.com*

Biological pathways are central to understanding complex diseases such as cancer. The majority of this knowledge is scattered in the vast and rapidly growing research literature. To automate knowledge extraction, machine learning approaches typically require annotated examples, which are expensive and time-consuming to acquire. Recently, there has been increasing interest in leveraging databases for distant supervision in knowledge extraction, but existing applications focus almost exclusively on newswire domains. In this paper, we present the first attempt to formulate the distant supervision problem for pathway extraction and apply a state-of-the-art method to extracting pathway interactions from PubMed abstracts. Experiments show that distant supervision can effectively compensate for the lack of annotation, attaining an accuracy approaching supervised results. From 22 million PubMed abstracts, we extracted 1.5 million pathway interactions at a precision of 25%. More than 10% of interactions are mentioned in the context of one or more cancer types, analysis of which yields interesting insights.

Keywords: Distant supervision, knowledge extraction, cancer pathways, Literome

1. Introduction

Cancer stems from the synergistic perturbation of multiple pathways by mutations.[1] Recent advances in sequencing technology offer a plethora of panomics data, holding the promise to make precision medicine and personalized treatment a reality. However, it remains a formidable challenge to identify cancer drivers, due to complex cross talks and feedback loops in cancer pathways.[2] Moreover, even when the drivers are identified, they may not be directly druggable, as in the case of RAS, where the promising targets lie in downstream signaling pathways.[3] Pathways are thus essential to understanding cancer and developing targeted treatments. As a result, pathways have been increasingly applied to panomics analysis.[4–8]

The majority of pathway knowledge resides in free text such as journal articles, which has been undergoing its own exponential growth. For example, PubMed contains over 22 million papers and adds more than one million each year. It is hard for manual curation to keep pace with such a vast and rapidly growing literature, making it a priority to automate the curation process. Such automation was traditionally pursued via rule-based systems,[9] but hand-coding extraction rules is expensive and time-consuming, and generally suffers low recall due to the varieties of ways for expressing the same meaning. Machine learning approaches offer a much more attractive alternative by effectively automating the rule engineering itself, but they in turn require annotated examples, which are still difficult to acquire in scale.

The lack of annotated examples can be compensated for by leveraging *distant supervision* from existing knowledge bases, as first proposed by Craven & Kumlien[10] and recently pursued actively in the natural language processing (NLP) community.[11–13] However, no existing approach addresses pathway extraction, focusing instead almost exclusively on newswire.

In this paper, we present the first attempt to apply distant supervision in pathway extrac-

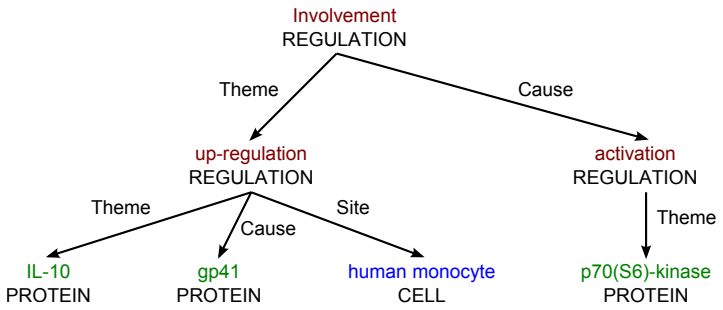

Fig. 1. Example pathway annotation of the sentence "Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein".

tion. We formulate pathway extraction as a classification problem and apply a state-of-the-art distant-supervision method to it. To better evaluate the effectiveness of distant supervision in bridging the gap from supervised learning, we propose a novel evaluation methodology to create a controlled experimental setting using the GENIA event extraction dataset.[14] Experimental results show that distant supervision outperforms baseline systems such as rule-based extraction by a wide margin, attaining an accuracy approaching supervised learning. Finally, we applied distant supervision to all PubMed abstracts, using prior pathway knowledge from the Pathway Interaction Database (PID).[15] Our system extracted 1.5 million pathway interactions at a precision of 25%. More than 10% of these interactions are cancer-related, analysis of which yields a number of interesting observations.

2. Methods

2.1. *Pathway Extraction from Text*

Biological pathways capture the interactions among genes, gene products, and small molecules such as metabolites. Examples of interactions include transcriptional regulation (e.g., transcription factor binding for transcription initiation) and post-translational regulation (e.g., kinase phosphorylation for protein activity modulation). For simplicity, in this paper we focus on static pathways such as signaling transduction and gene regulation, rather than metabolic networks and dynamics. Figure 1 shows an example pathway fragment, with the corresponding text and annotation. In general, pathways form a hypergraph where each node represents a gene or gene product, and each hyperedge represents an interaction.

Decades of genetic studies have produced a wealth of pathway knowledge, which is scattered in the literature, with each paper or sentence covering only a few interactions. For example, the sentence "CTCF is a transcriptional repressor of the c-myc gene." signifies a transcriptional regulation of c-myc by CTCF. In this paper, we focus on extracting such regulatory relations between two proteins, which identify the `Cause` argument such as CTCF, the `Theme` argument such as c-myc, and, if available, the regulation direction such as `negative_regulation`. Formally, the pathway extraction problem is to classify each ordered triple (P_T, S, P_C) into one of the following: {`positive_regulation, regulation, negative_regulation, NULL`},

Fig. 2. A simple distant supervision algorithm

Require: A set of sentences, with entity mentions identified
Require: A database of relation triples (entity, relation, entity)
1: For each relation triple, find all sentences containing the entity pair
2: Annotate those sentences with the corresponding relation
3: Sample unannotated sentences with co-occuring proteins as negative examples
4: Train a classifier using the annotated dataset
5: **return** the resulting classifier

where S is a sentence and P_T, P_C are protein mentions in S.*

2.2. *Distant Supervision*

The main idea of distant supervision is to use known relation instances in the database to automatically annotate training examples in unlabelled text,[10] as shown in Figure 2. Suppose we know from the database that CTCF down-regulates c-myc, and in the text we find a sentence where CTCF and c-myc co-occur, such as "CTCF is a transcriptional repressor of the c-myc gene". We thus have some reason to hypothesize that this sentence might be stating a `negative_regulation` relation with `Theme` being c-myc and `Cause` being CTCF. A simple distant-supervision method would thus label such sentences as positive examples, and sample negative examples from random sentences where the co-occurring proteins do not have a known regulatory relation.

Of course, this simple approach would often introduce noise in the labels, since the sentence might not be about the given regulation, as in "In Bcl-deficient mice, expressions of both CTCF and c-myc showed marked decrease". A more reasonable assumption is that *some* sentence in the text expresses this relation, though not necessarily every one with co-occuring CTCF, c-myc.† This assumption is adopted in state-of-the-art distant supervision methods.[12,13]

In this paper, we use MultiR[13] because it is such a state-of-the-art method with a publicly available implementation. The key idea is to introduce a latent variable to signify whether a relation R holds between entities (E_1, E_2) for each sentence where E_1 and E_2 co-occur and are the `Theme` and `Cause`, respectively. Distant supervision is provided by enforcing during training that for each relational triple (E_1, R, E_2), at least one latent assignment is true if the database contains the relation, and none otherwise. Each instance is represented by a linear model with features over the sentence and entities. The feature weights are learned using online learning with perceptron. In each iteration, for each protein pair, MultiR first computes the best assignment to each instance according to the model. If the assignment is consistent with the database (each relation is expressed at least once, and no relations that don't appear in the database are expressed), no update is done on the protein pair. If not, it

*This formulation might sometimes lose information, such as co-factors required in a regulation, or experimental conditions. Lifting this limitation to handle n-ary, nested relations will be a key future direction.
†Note that this assumption still suffers a number of drawbacks. For example, it is possible that none of the available sentences mention the given relation. Moreover, the existing database is incomplete, so the absence of a relation might not necessarily signify its negation. Addressing these issues is an active research area.

uses a greedy algorithm to find the best assignment to each instance such that the assignment is consistent with the database, and does a perceptron update toward this assignment.

We used the following standard lexical and syntactic features[11] in our experiments, illustrated with sentence "Involvement of p70(S6)-kinase activiation in IL-10 up-regulation in human monocytes by gp41 envelop protein", and E_1, E_2 being "gp41" and "p70(S6)-kinase", respectively.

Direction: 0 if the two protein spans overlap, +1 if E_2 follows E_1 in the sentence, and −1 if E_2 precedes E_1. In the example, this feature is -1.

Distance: Four indicator features specifying whether E_1 is more than k tokens to the right of E_2, with $k = 5, 10, 15, 20$. In the example, this feature is 1 for $k = 5$ and 0 otherwise.

Lexical: (1) the sequence of words between the two proteins, concatenated with direction, such as `Dir=-1` $\wedge$ `WordSeq="activation in IL-10 up-regulation in human monocytes by"`; (2) the sequence of lemmas of the words between the two proteins, concatenated with direction, such as `Dir=-1` $\wedge$ `LemmaSeq="activate in il-10 up-regulate in human monocyte by"`; (3) individual words between the two proteins, concatenated with direction, such as `Dir=-1` $\wedge$ `Word="activation"`, etc.; (4) individual lemmas between the two proteins, concatenated with direction, such as `Dir=-1` $\wedge$ `Lemma="activate"`, etc.

Dependency path: (1) unlexicalized dependency path between the two proteins (e.g. $\uparrow_{nn}\uparrow_{by}\uparrow_{in}\downarrow_{of}\downarrow_{nn}$); (2) lexicalized dependency path using the lemmas for lexicalization (e.g. $\uparrow_{nn}$ *protein* $\uparrow_{by}$ $up-regulate$ $\uparrow_{in}$ *involve* $\downarrow_{of}$ *activate* $\downarrow_{nn}$); (3) upward (child to parent) and downward (parent to child) portions of the dependency path as separate features, concatenated with the lemma of the path root, e.g., `Upward=`$\uparrow_{nn}\uparrow_{by}\uparrow_{in}$ $\wedge$ `RootLemma="involve"` and `Downward=`$\downarrow_{of}\downarrow_{nn}$ $\wedge$ `RootLemma="involve"`.

MULTIR can be used in supervised learning by simply setting each relational assignment according to the sentence-level annotation (i.e., they are no longer latent).

2.3. *Controlled Experiments using GENIA Event Extraction Dataset*

Evaluating distant supervision is challenging as by definition there is no annotated dataset, so existing methods tend to resort to reporting sample precision and absolute recall (i.e., sample a small subset of system extractions, manually inspect them to determine the precision, and use it to estimate the number of correctly extracted instances). While this is useful for comparing distant supervision methods, the sampling process inevitably introduces bias and variance. Furthermore, it is difficult to assess the performance gap from supervised learning.

This motivates us to propose a new evaluation methodology by creating a simulated distant-supervision scenario from an annotated dataset, which enables us to assess the true precision and recall, and compare with supervised learning. Specifically, we used the GENIA event extraction dataset from BioNLP-09 Shared Task 1,[14] where protein annotation is given as input, and pathway events are annotated as output (Figure 1).

We follow the formulation in Section 2.1 and reduce GENIA events to binary rela-

tions in { `positive_regulation, regulation, negative_regulation, NULL` }.[‡] Specifically, for each protein pair E_1, E_2 in a training sentence, we compute all event paths between them from the annotation, and reduce each event path into a relation summarizing the path semantics as follows:

First, we identify the top event `e` in the path that has both proteins in its scope. E_1 should lie in the `Theme` branch from `e`, and E_2 in the `Cause` branch. If not, the relation is set to `NULL`.

Next, we check whether the `Theme` path contains any `Cause` argument, as between TP53 and MAPK1 in "TP53-induced BCL overexpression is inhibited by MARK1". If so, the relation is also set to `NULL`. In general, as we can see from this example, we can not conclude a causal relation between these two proteins.

Otherwise, we assign a regulation relation with E_1 being the `Theme` argument and E_2 being the `Cause` argument. If any event in the path is `regulation`, we set the overall relation to `regulation` as well (i.e., we can not determine the direction). Otherwise, we set the relation to `positive_regulation` if there are an even number of `negative_regulation` events followed by a `Theme` argument, and to `negative_regulation` for an odd number.

For example, with sentence "Involvement of p70(S6)-kinase activiation in IL-10 up-regulation in human monocytes by gp41 envelop protein", and E_1, E_2 being "gp41" and "p70(S6)-kinase", respectively, the non-NULL relations are `(positive_regulation, IL-10, gp41), (regulation, IL-10, p70(S6)-kinase)`.

We thus form a database with the unique relation triples derived from the training set, which, together with the unlabelled text are served as input to the distant supervision method. The learned classifier is then applied to the test text to extract new events. The results can be evaluated using the test annotation and directly compared with the supervised learning approach, which is separately learned using sentence-level annotation in training text.

2.4. *Rule-based Relation Extraction*

In the past, rule-based extraction has been used extensively in relation extraction, such as the GeneWays system.[9] This remains a preferred approach in the absence of annotated examples for applying machine learning approaches.[16] To enable a head-to-head comparison with distant supervision, we followed standard practice as in GeneWays and other rule-based systems, and spent substantial effort to develop a high-performing rule-based system for pathway extraction.

Upon first consideration, relation extraction appears straightforward to automate. For example, in the sentence "CTCF is a transcriptional repressor of the c-myc gene.", "transcriptional repressor" clearly indicates a `negative_regulation` relation. Unfortunately, the same information may be expressed in many variations, as shown in Table 1.

Some variations can be normalized in syntactic analysis. For example, lemmatization can normalize inflectional morphology ("inhibited" has the lemma "inhibit"), and derivational morphology (both "inhibitor" and "inhibition" are derived from "inhibit"). Syntactic parsing can normalize active/passive variations and identify related words even when they are far

[‡]GENIA also annotates simple unary events such as `expression` and `transcription`, which could be taken into account in future work.

Table 1. Linguistic varitions describing the same pathway event.

Sentence	Varation
CTCF is an inhibitor of the c-myc gene	lexical
CTCF inhibits the c-myc gene	part of speech
The c-myc gene is inhibited by CTCF	active/passive voice
The ability of CTCF to inhibit c-myc	modality
CTCF has been shown to inhibit c-myc	background
Expression of the CTCF gene has been shown to inhibit c-myc	explanation
CTCF as well as other genes have been shown to inhibit c-myc	augmentation

apart in the sentence.

Other variations, however, are more difficult to handle in a domain-independent way, so past systems resorted to writing domain-specific rules to capture such variations.[9,16] We followed this approach and developed our rules based on the event annotation in the training set of GENIA. The rules identify subtrees that trigger a particular relation, as well as child subtrees that identify the `Cause` and `Theme` of that relation. At the surface level, one might say that "`Cause` inhibits `Theme`" is a trigger for `negative_regulation`. We represent these rules in terms of syntactic subtrees to better handle many of the aforementioned variations: "(inhibit nsubj: `Cause` dobj: `Theme`)" triggers `negative_regulation`.

To identify relations for a candidate sentence and protein pair, we first parse the sentence using SPLAT[17] and postprocess the parse into Stanford typed dependencies,[18] which forms a tree where each node represents a word by its surface form, its lemma, and its part of speech, and each edge represents a typed dependency. For example, for "CTCF inhibits c-cmyc", "inhibits" is the root, with a "`subject`" being "CTCF" and an "`object`" being "c-myc". Next we attempt to match each trigger rule at each node in the tree. Here, matching means finding a correspondence between the nodes in the trigger subtree and the nodes in the candidate sentence. Each trigger subtree node must map to a unique sentence node, with matching relation types and node lemmas, except for `Cause` and `Theme`, which can match any nodes in the input tree. Finally, we check if the `Cause` and `Theme` are compatible with the protein arguments. In the simplest case, this amounts to matching `Cause` and `Theme` to the corresponding protein nodes. Additionally, we expand the criteria to account for variations in protein mentions (e.g., matching to "gene" in "the BCL2 gene", or matching to "BCL1" in "BCL1 and BCL2"). If this is successful, we return the specified relation as the classification; otherwise, `NULL`.

This results in a set of 159 trigger rules and 63 protein expansion criteria, by hill-climbing on extraction accuracy over the GENIA training set.

2.5. *PubMed Extraction with Distant Supervision from PID*

Given the GENIA dataset, one might consider simply adopting supervised learning and applying the learned extractor to PubMed abstracts, as in EVEX.[19] Such extractions would no doubt be very useful, but there remains a major concern over how representative the examples are, as with any supervised approach. The GENIA abstracts were chosen a decade ago and are rather dated by now. More importantly, they were sampled from a narrow subject area

(PubMed search with MeSH terms "human", "blood cell" and "transcription factor").

In contrast, a distant supervision approach can learn an extractor for any subject area, as long as there exist manually curated databases to leverage, which is generally the case. To demonstrate the feasibility of this direction, we used PID[15] as the database for distant supervision, and applied the learned extractor to PubMed-scale pathway extraction.

PID represents each pathway as a hypergraph, where each node represents a gene (transcription), gene product (proteins and complexes), or process (e.g., apoptosis), and each hyperedge represents an interaction, with multiple input and output nodes and their regulatory directions (induction or inhibition). We followed the formulation in Section 2.1 and reduced each interaction into binary regulations by creating relation triples between the component genes in each input/output pair, excluding the case when the two nodes represent identical molecules. This yields 15150 unique relation triples, such as (`negative_regulation, APC, TP53`). We filter out triples with conflicting regulation or causal directions between two proteins, such as (`positive_regulation, CDKN1A, TP53`), (`negative_regulation, CDKN1A, TP53`), and (`positive_regulation, PLAU, PLG`), (`positive_regulation, PLG, PLAU`). These are legitimate interactions representing feedback loops or contextual dependency, but their inclusion would confuse the distant-supervision learner given the lack of sentence-level annotation. There were 4,547 triples left after filtering, which were used for distant supervision.

In GENIA, gold protein annotation is given as input, which is not available for general PubMed abstracts. We thus used the protein extractor from Literome[20] to identify protein mentions in all PubMed abstracts. This extractor was built on various available resources with canonical mentions and synonyms for proteins, families, and complexes.

General PubMed abstracts are significantly more diverse and noisy compared to GENIA ones. At training time, we filtered out instances where the two proteins have overlapping spans or appear more than 15 words apart, which are unlikely to express a relation.

2.6. *Cancer Classification*

Given a pathway interaction, it would be useful to understand the context such as cell type, localization, experimental conditions, etc. As a first step toward this direction, we focus on identifying the cancer types mentioned in the same abstract. Specifically, we used the MeSH terms provided by Medline for the majority of PubMed abstracts, and extracted the ones that signify cancer types (i.e., ending in "Neoplasms"). An extraction instance is thus associated with the cancer types for the given abstract.

3. Results

3.1. *GENIA Experiments*

We used the GENIA dataset to create a controlled setting to evaluate distant supervision and compare it with rule-based and supervised approaches (Section 2.3). The GENIA dataset[14] contains a set of annotated abstracts (800 training, 150 development). It also contains a test set, but its annotation is not made public so we can not use it to evaluate binary relation extraction. Therefore, we conducted training and development using the training data, and

Table 2. Test results on GENIA binary-relation classification comparing distant supervision with two baseline systems, supervised learning, and MSR11, a state-of-the-art system training on full event structures.

System	Precision	Recall	F1
Most-Frequent	3.4	69.7	6.5
Rule-Based	45.8	5.2	9.4
Distant Supervision	39.2	19.0	25.6
Supervised	37.5	29.9	33.2
MSR11	55.1	28.0	37.1

reserved the development set for test. We subsampled negative examples to avoid label imbalance (the ratio between positive and negative examples is about 1:3). For supervised learning, we also filtered out features with fewer than 3 counts in positive examples. Training and test both took less than a second.

We took all co-occuring protein pairs in test sentences as classification candidates and evaluated precision, recall and F1 of system extraction given the gold labels. We compared distant supervision with the rule-based system, a baseline that predicts the most frequent relation in training (`positive_regulation`) for co-occurring protein pairs, as well as the supervised system trained on sentence-level annotation, which provides an upper bound. To quantify a further upper bound when full event structures are taken into account, we also evaluated MSR11[21] by converting its publicly available event extraction output to binary relations. MSR11 is an event extraction system trained on full event structures, with state-of-the-art event F1 score of 55.7 on GENIA development. (The best score of 58.7 is reported by Riedel et al.[22]) Note that event F1 accounts for simple unary events such as `expression`, thus is not directly comparable with binary-relation F1. For example, MSR11 scores only 37.1 on binary-relation F1.

Table 2 shows the test results for all systems. Surprisingly, not only did distant supervision substantially outperform the two baseline systems, it also attained an accuracy that is rather close to the supervised upper bound. For example, compared to the rule-based system, distant supervision reduces the performance gap from the supervised upper bound by nearly 70%, while using much less information. (The rule-based system used sentence-level annotation during rule engineering.) Note that this supervised upper bound is very strong, comparable to state-of-the-art event extraction that leverages full event structures (33.2 vs. 37.1). Likewise, we spent substantial effort to engineer the rule-based system, attaining a strong performance on training (precision 80.3, recall 26.5, F1 39.8). Unfortunately, while its precision is relatively high, it suffers low recall and a big performance drop from training to test, which is typical of rule-based systems. Overall, these results clearly demonstrate the promise of distant supervision, which could attain competitive accuracy while requiring substantially less development effort compared to both rule-based and supervised approaches.

3.2. *PubMed Experiments*

We applied distant supervision using PID and PubMed abstracts to train an extractor (Section 2.5). We sampled negative examples following the positive/negative ratio used in GENIA, yielding 97,215 examples in total. We then ran the extractor to extract events from all PubMed

Table 3. Evaluation on 300 sample PubMed extractions, with annotation statistics and example instances. Mentions are annotated with `Cause` and `Theme` from automatic predictions.

Outcome	**Count**	**Example**
Correct	75	The polycomb protein Cause[Bmi-1] represses the INK4a locus , which encodes the tumor suppressors Theme[p16 and p14(ARF)].
Imperfect Sign	27	This regulated control of Theme[STAM] expression by Cause[Hrs] was independent of transcription.
Reversed Direction	46	This may possibly occur through inhibition of Cause[insulin receptor (IR)] tyrosine kinase activity mediated by serine/threonine phosphorylation of the IR or Theme[insulin receptor substrate 1 (IRS-1)].
Protein Error	56	We found that the development of experimental autoimmune encephalomyelitis (EAE), the rodent model of Cause[multiple sclerosis], was significantly suppressed in Theme[IL-17] (-/-)-mice.
Non-Regulation	96	Anti-dsDNA B cells, on the other hand, are functionally unresponsive to anti-IgM and Theme[LPS] stimulation, and do not phosphorylate intracellular proteins, including Cause[Syk], upon mIg stimulation.

abstracts by classifying candidate sentences with co-occurring protein pairs. Note that we do not know the gold annotation for the positive examples used in training, nor could the learner memorize them since no protein-specific features are in use. Training took five minutes and extraction took 30 minutes using 900 cores. This PubMed-scale extraction yields 1,491,373 regulation instances, with 838,255 unique relation triples.

To assess the quality of the extraction, we sampled 300 extractions and manually annotated them. Table 3 summarizes our findings. Among the 300 sample extractions, 56 have wrong protein annotation, which is not surprising given that protein mentions are often highly ambiguous. Of the remaining 244 instances, 75 are correct, giving an end-to-end precision of 25% and a precision of 31% assuming gold protein annotation. Among the errors, 46 are actually correct regulation events, but in 21 of them the sign (`positive_regulation` or `negative_regulation`) is wrong or the sentence is ambiguous about it, and in the remaining 25 the causality direction is reversed. With this sample precision, we estimate that distant supervision from PID yields about 372,000 correct extractions, and 210,000 unique relation triples, which is an order of magnitude larger than PID.

These results are promising and testify to the feasibility of this direction. Of course, there is still much room for improvement. Protein errors occur in about one fifth of extractions. Currently, protein extraction does not benefit from distant supervision and is done separately from event extraction. Joint learning of protein and event extraction with distant supervision could potentially produce large improvement in both tasks. Relation errors often occur for two proteins in paths that are in conjunction, as in the example in Table 3, which might potentially be avoided with better filtering criteria. Finally, distant supervision can be used in

Table 4. Top ten most studied cancer types, along with the top ten genes for each type, both in the number of unique pathway extractions.

Cancer	Relations	Top ten most studied genes
Breast	27988	TP53, ESR1, MYBL2, BRCA1, ZFP36, EGFR, ZFP, ESR2, EGF, AKT
Prostate	10981	CBX8, SLC22A3, AR, KLK3, EPHB2, TP53, TDRD7, NPEPPS, AKT, SERPINB6
Lung	9423	EGFR, TP53, KRAS, VEGFA, CASP, MMP2, AKT, EPHB2, CDH1, CEACAM5
Liver	8438	TP53, EPHB6, HCCS, AFP, MYLIP, CCL2, VEGFA, RELA, NFKB1, NA
Colon	6092	TP53, APC, CTNNB1, AOM, TMED7, RELA, EGFG, NFKB1, SRC, PTGS2
Colorectal	5381	TP53, CTNNB1, APC, CEACAM5, KRAS, EGFR, MSI, CASP, VEGFA, PTGS2
Pancreatic	5178	INS, KRAS, EGFR, VEGFA, CEACAM5, TNF, AURKA, MIA, RELA, NFKB1
Ovarian	4331	LPA, TP53, EGFR, VEGFA, AKT, MMP, MUC16, BRCA1, BARX2, IFNG
Skin	4324	SERPINB3, TP53, KRT13, NFKB1, RELA, CD4, VIM, TNF, IL2, CD68
Brain	3988	RIPK1, CCDC88A, TP53, CSF2, NCAM1, AFP, MGMT, EGFR, ELAVL1, AKT

Table 5. Top ten most studied cancer types, along with top genes ranked by association score.

Cancer	Top ten most significantly associated genes
Breast	ESR1, BRCA1, ESR2, BRCA2, PGR, CYP19A1, ERBB2, EGF, IGF1, KRAS
Prostate	CBX8, KLK3, NPEPPS, AR, DYNLL1, SERPINB6, ERG, DPT, TMPRSS2, FOLH1
Lung	EGFR, KRAS, ALK, PCSK9, CBX8, ARCN1, KLK3, BRCA1, F7, TTF1
Liver	HCCS, EPHB6, AFP, TRIM26, HSPG2, ADAM17, LDLR, DNLZ, ALB, CCL15
Colon	AOM, DLD, CEACAM5, DDX53, APC, CTNNB1, GAST, PPARG, WNT16, SELE
Colorectal	KRAS, APC, MSI2, CTNNB1, MSI1, CEACAM5, MRC1, BRAF, FAP, MLH1
Pancreatic	INS, GCG, MIA, SST, CCK, KRAS, ZGLP1, PDX1, PRSS27, SMAD4
Ovarian	BRCA2, MUC16, BRCA1, LPA, DIRAS3, BRCA3, ARID1A, HEY1, GNRHR, ABCB1
Skin	SERPINB3, TNFRSF8, COL1A1, CMM, MLANA, EGFR, MC1R, CPD, CD8A, MCC
Brain	GFAP, MGMT, IDH1, MS, SMS, NEFH, KIAA1549, CSF2, GSC, NAA60

an integrative loop for eCuration:[23] distant supervision produces initial extractions, eCurators then verify them via an online interface such as Literome,[20] fixing errors by a click of buttons, which is much more efficient than annotation from scratch. The feedback could be fed back into distant supervision via online learning and used to continuously improve extraction quality. Active learning can also be incorporated by prioritizing the least confident extractions for annotator verification.

3.3. *Cancer Pathway Analysis*

With cancer contexts identified by MeSH terms (Section 2.6) and PubMed extractions produced from distant supervision (Section 3.2), we conducted an exploratory analysis to survey the research landscape and findings on cancer pathways.

Cancer Pathway Research Among the 1.5 million pathway extractions, 150,379 occur in the context of cancer, or about 10%; among the 838 thousand unique pathway relations, 108,373 occur in the context of cancer, or about 13%. Table 4 shows the top ten most studied cancer types, along with the top ten genes for each type, both in the number of unique pathway relations found in our extraction. Not surprisingly, many well-known cancer genes are in the list, such as TP53 and EGFR. Overall, the top ten most studied genes for cancers are: TP53, EGFR, VEGFA, CBX8, ESR1, SLC22A3, AKT, MYLIP, EGF, EPHB2. For non-cancer context: INS, TNF, CA2, TCF, CD4, TP53, IRF6, EPHB2, CALM3, CASP.

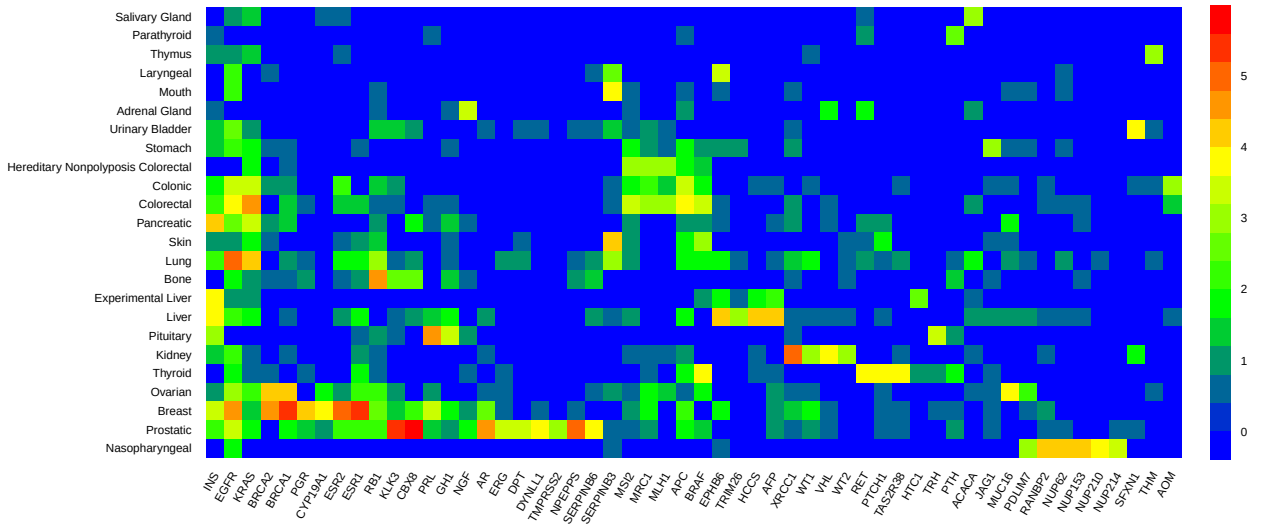

Fig. 3. Heatmap indicating the strength of association between cancer types and genes.

Cancer-Specific Genes Genes like TP53 are crucial for many cancer types and are well studied across the board. Additionally, we are interested in genes that act specifically to a cancer type. We thus searched for non-random association between genes and cancer types, using the log-likelihood ratio test where the null model assumes that the gene occurs independently from the cancer type in each extraction. Figure 3 shows a heatmap for the association scores between the top ten cancers and their most significantly associated genes. Table 5 shows the top ten most studied cancer types again, along with the top ten genes ranked by association score instead. Compared to top genes by relation counts, these lists are clearly type-specific, revealing many well-known associations, such as BRCA1, BRCA2, ESR1, ESR2 for Breast Cancer, EGFR, KRAS for Lung Cancer, BRCA1, BRCA2, BRCA3 for Ovarian Cancer, etc. Conversely, the highly ranked genes with lower extraction counts might reveal research opportunities for genes that are not yet well studied but potentially important for a cancer type. In general, our PubMed-scale extraction enables cancer pathway analysis encompassing the entire research landscape, revealing interesting insights as well as suggesting future research priorities.

4. Discussion

In this paper, we present the first attempt to apply distant supervision to pathway extraction. Evaluation on the GENIA event extraction dataset shows that distant supervision substantially outperforms rule-based extraction and other baselines, attaining an accuracy approaching supervised upper bounds. Application to all PubMed abstracts using the PID database yielded an order of magnitude more correct pathway interactions than the original database. Analysis of cancer-related interactions led to a number of interesting observations. The extracted pathways, sample annotation, and trigger rules used by the rule-based system will be made available at `literome.azurewebsites.net/papers/psb15`.

Overall, this demonstrates the great potential of distant supervision for cancer pathway

extraction and biological knowledge extraction in general: distant supervision could attain high accuracy while requiring substantially less development effort compared to both rule-based and supervised approaches.

This also opens up a number of interesting future research directions. Currently, pathway extraction is pursued in an isolated fashion, feeding on output from other tasks such as protein extraction. Joint learning with distant supervision is a promising direction for improving the accuracy in all pipeline tasks. Existing distant supervision methods are only applicable to binary relations; lifting this limitation to handle n-ary and nested relations is an important future direction, and is particularly important for identifying relevant contexts and reconciling seemingly conflicting relations. Our system currently ignores event modalities such as negation and hedging, which need to be incorporated in the future. Distant supervision can be seamlessly integrated in eCuration, combining with online learning from annotator feedback, and active learning for prioritizing verification requests. Finally, a particularly exciting prospect is to integrate the extracted pathways with high-throughput panomics data for automating discovery in genomic medicine.[5]

References

1. D. Hanahan and R. A. Weinberg, *Cell* **144(5)**, 646 (2011).
2. B. Vogelstein, *et al.*, *Science* **339** (2013).
3. A. G. Stephen, *et al.*, *Cancer Cell* **25(3)**, 272 (2014).
4. K. Wang, *et al.*, *Nature Reviews Genetics* **11**, 843 (2010).
5. C. J. Vaske, *et al.*, *Bioinformatics* **26**, 237 (2010).
6. T. Ideker, *et al.*, *Cell* **144**, 860 (2011).
7. S. Ng, *et al.*, *Bioinformatics* **28**, 640 (2012).
8. A. J. Sedgewick, *et al.*, *Bioinformatics* **29**, 62 (2013).
9. A. Rzhetsky, *et al.*, *J Biomed Inform.* **37(1)**, 43 (2004).
10. M. Craven and J. Kumlien, Constructing biological knowledge bases by extracting information from text sources, in *Proc. of the 7th International Conference on Intelligent Systems for Molecular Biology*, 1999.
11. M. Mintz, *et al.*, Distant supervision for relation extraction without labeled data, in *Proc. of the Forty Seventh Annual Meeting of the Association for Computational Linguistics*, 2009.
12. S. Riedel, *et al.*, Modeling relations and their mentions without labeled text, in *Proc. of the Sixteen European Conference on Machine Learning*, 2010.
13. R. Hoffmann, *et al.*, Knowledge-based weak supervision for information extraction of overlapping relations, in *Proc. of the Forty Ninth Annual Meeting of the Association for Computational Linguistics*, 2011.
14. J.-D. Kim, *et al.*, Overview of BioNLP-09 Shared Task on event extraction, in *Proc. of the BioNLP Workshop*, 2009.
15. C. F. Schaefer, *et al.*, *Nucleic Acids Research* **37**, 674 (2009).
16. K. Ravikumar and H. Liu, Towards pathway curation through literature mining - a case study using PharmGKB, in *Proc. Pacific Symposium of Biology*, 2014.
17. C. Quirk, P. Choudhury, J. Gao, H. Suzuki, K. Toutanova, M. Gamon, W. Yih and L. Vanderwende, MSR SPLAT, a language analysis toolkit, in *Proc. of NAACL HLT Demonstration Session*, 2012.
18. M.-C. de Marneffe, *et al.*, Generating typed dependency parses from phrase structure parses, in *Proc. of the Fifth International Conference on Language Resources and Evaluation*, 2006.
19. S. V. Landeghem, *et al.*, *PLoS One* **8** (2013).
20. H. Poon, *et al.*, *Bioinformatics* (2014).
21. C. Quirk, *et al.*, MSR-NLP entry in BioNLP Shared Task 2011, in *Proc. BioNLP*, 2011.
22. S. Riedel and A. McCallum, Fast and robust joint models for biomedical event extraction, in *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 2011.
23. C.-H. Wei, *et al.*, *Database* (2012).