BioRelEx 1.0: Biological Relation Extraction Benchmark

Hrant Khachatrian[1,2], Lilit Nersisyan[3], Karen Hambardzumyan[1,2], Tigran Galstyan[1,2],
Anna Hakobyan[3], Arsen Arakelyan[3], Andrey Rzhetsky[4], and Aram Galstyan[5]

[1]YerevaNN, Yerevan, Armenia
[2]Department of Informatics and Applied Mathematics, Yerevan State University, Yerevan, Armenia
[3]Bioinformatics Group, Institute of Molecular Biology, NAS RA, Yerevan, Armenia
[4]Institute for Genomics and Systems Biology, Departments of Medicine and
Human Genetics, University of Chicago, Chicago, Illinois, USA
[5]USC Information Sciences Institute, Marina Del Rey, California, USA
hrant@yerevann.com

Abstract

Automatic extraction of relations and interactions between biological entities from scientific literature remains an extremely challenging problem in biomedical information extraction and natural language processing in general. One of the reasons for slow progress is the relative scarcity of standardized and publicly available benchmarks. In this paper we introduce BioRelEx, a new dataset of fully annotated sentences from biomedical literature that capture *binding* interactions between proteins and/or biomolecules. To foster reproducible research on the interaction extraction task, we define a precise and transparent evaluation process, tools for error analysis and significance tests. Finally, we conduct extensive experiments to evaluate several baselines, including SciIE, a recently introduced neural multi-task architecture that has demonstrated state-of-the-art performance on several tasks.

1 Introduction

Biological interaction databases capture a small portion of knowledge depicted in biomedical papers, due to time consuming nature of manual information extraction. As experimental methodologies to identify such interactions tend to increase in scale and throughput, the problem stands to rapidly update these databases for relevant applications (Oughtred et al., 2018). The long-term aim of our efforts is to provide bases for filling this gap automatically.

Despite significant progress in recent years, extracting relationships and interactions between different biological entities is still an extremely chellenging problem. Some of those challenges are due to objective reasons such as lack of very large annotated datasets for training complex models, or wide variability in biomedical literature which can lead to domain mismatch and poor generalization. Another important challenge, which is the main focus of the present paper, is the scarcity of publicly available datasets. Indeed, with despite some notable exceptions (Kim et al., 2003; Dogan et al., 2017), there is a relative lack of adequate, high-quality benchmark datasets which would facilitate reproducible research and allow for robust comparative evaluation of existing approaches.

Here we have processed biological texts to annotate biological entities and interaction pairs. In contrast to other related databases, our efforts were focused on delineation of biological entities from experimental ones, and on distinguishing between indirect regulatory interactions and direct physical interactions. Furthermore, we have performed grounding via cross-reference of annotated entities with external databases. This allows for merging interactions from different sources into a single network of biomolecular interactions.

The main contributions of this work are:

1. We publish a dataset of 2010 sentences with complete annotations of biological entities and binding interactions between the entities,

2. We propose a benchmark task with a well-defined evaluation system, which follows the best practices of machine learning research,

3. We perform extensive evaluation of several competing methods on the dataset and report the results.

Proceedings of the BioNLP 2019 workshop, pages 176–190
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

2 Related work

In this section we briefly summarize prior work on relation extraction from unstructured text.

Since 2009, NIST has organized Knowledge Base Population evaluations as part of Text Analysis Conferences (TAC KPB). Thousands of sentences from newswire and informal web pages were annotated for training and evaluation purposes (Getman et al., 2018). In 2017, a team from Stanford released TACRED (Zhang et al., 2017), a dataset of 106 264 sentences with 42 relation types. The relations are mainly between people, places and organizations.

A large number of papers focused on biological relation extraction. (Bunescu et al., 2005) built a manually annotated corpus of 225 abstracts to evaluate various extraction methods. This dataset is referred as AIMed in subsequent papers. Later, (Pyysalo et al., 2007) developed a smaller dataset called BioInfer with more detailed annotations. In particular, the authors developed large ontologies for biological entities and relations between them and attempted to classify each entity and relation according to these ontologies. The small number of sentences and interactions is 1100 and 2662, respectively, so for many types of relations there were too few samples. Because of that, almost all subsequent papers that applied machine learning techniques on BioInfer discarded the detailed labels and used it as a dataset of binary relations. In 2008, (Pyysalo et al., 2008a) presented a detailed comparison of AIMed, BioInfer and three other datasets (IEPA, HPRD50 and LLL) and found significant differences in the data collection and evaluation procedures.

In (Pyysalo et al., 2008b), the authors concluded that the results on the five datasets reported in different papers are incomparable and suggested to unify the datasets in a common format with a precise evaluation procedure. This proved to be successful as a large number of subsequent papers use the unified versions of the datasets. On the other hand, these datasets are currently used only for binary relation classification, as the unified versions keep the lowest common level of annotations (only entity locations and binary labels between the pairs). It means that the models trained on these datasets cannot be used for end-to-end relation extraction from text. Moreover, many recent papers violate evaluation strategies (e.g. perform cross-validation on splits that do not respect document boundaries) and report unrealistically high scores (Hsieh et al., 2017; Ahmed et al., 2019).

One of the highest quality datasets is developed as part of GENIA project (Kim et al., 2003). It involves annotations of entities, syntactic features, wide variety of events, including around 2500 binding interactions (Thompson et al., 2017). GENIA does not have a training/test split, but various subsets of it have been used as training and test sets of BioNLP Shared Tasks in 2009 (Kim et al., 2009), 2011 (Kim et al., 2011) and 2013 (Nédellec et al., 2013). Several protein-protein interaction (PPI) datasets appeared in BioCreative series of shared tasks. There was a track on PPI extraction in BioCreative II, including a binary relation extraction subtask from full texts and another subtask for finding evidence sentences for the given interaction (Krallinger et al., 2008). BioCreative V Track 4 included a subtask on extraction of more complex data structures called Biological Expression Language (BEL) statements (Rinaldi et al., 2016).

Other biological relation extraction datasets include ADE (Gurulingappa et al., 2012), a dataset of adverse drug effects; BB3 (Deléger et al., 2016), a dataset of relations between bacteria and their habitats, which was used in BioNLP Shared Task 2016; SeeDev, a dataset of sentences about seed development of plants; AGAC, a dataset on gene mutations and diseases. The latter three datasets are included in BioNLP 2019 Shared Tasks. Precision Medicine Track of BioCreative VI (Dogan et al., 2017) introduced a large dataset of protein-protein interactions that are affected by mutations.

SemEval 2017 Task 10 (Augenstein et al., 2017) was about extracting relations from scientific paper abstracts (physics, computer science and materials science). SemEval 2018 Task 7 focused on sentences from computational linguistics papers. SciERC (Luan et al., 2018) is a dataset consisting of 500 research paper abstracts from major AI conferences with annotated entities, coreference links and relations between entities.

3 Dataset description

3.1 The choice of sentences

We have annotated 2010 sentences for binding interactions between biological entities. Those sentences came from a much larger set of 40,000 sentences that were automatically extracted from various biomedical journals and underwent minimal manual post-processing (Rzhetsky et. al., 2019). While the original set contained numerous interaction types, here our focus is on binding interactions only. The text of the sentences are mostly copied from the journal websites and can include uncommon Unicode symbols. In rare cases we had to copy the sentences from PDF versions of the papers and manually fix incorrect characters.

As stated above, the current version of the dataset is focused on binding interactions. All sentences in the dataset contain one of the following words: "bind", "binds", "binding", "bound". This will potentially limit the applicability of the models trained on this dataset on other sentences that contain information about binding interactions.

3.2 Entities

3.2.1 Entity definition

Every annotated entity is a continuous span of characters in the sentence surrounded by non-alphanumeric symbols (can include spaces, hyphens etc.).

Tokenization of biomedical texts can be a challenging task. To ensure consistency, we have verified that all annotated entities in the dataset are surrounded by the symbols described in Table 3 of Appendix A.3. Note that all these symbols can also appear inside an entity name.

3.2.2 Entity types

We have annotated 33 types of entities. For classification of entities we were governed both, by biological function and by chemical structure. More specifically, we distinguish between biological and experimental entities. For example, if the sentence refers to an oligonucleotide in an experiment, we do not annotate it as DNA, but as an experimental-construct. Furthermore, we define main organic entity types as protein, protein-family, protein-complex, DNA and RNA, while refer to the rest of organic compounds as chemicals. The complete list of entity types is listed in Appendix A.4.

These decisions were motivated by two main reasons: (a) only biological entities should be annotated and cross-referenced in order to arrive at biologically meaningful interaction networks; (b) a higher level of annotation that disregards details (e.g. chemicals) significantly reduces annotation resources with no loss to our targetted aim. This contrasts to the Genia ontology, where entity annotation was only based on chemical structure of substances (Thompson et al., 2017).

Note that while the majority of entities are annotated to a single type, two entities with the same name may be annotated to different types (e.g. protein or protein-family) depending on the context, and sometimes these cases may co-occur in the same sentence (e.g. protein and gene (`1.0.train.166`)).

3.2.3 Coreference

Pairs of entities may be in *is_a* or *part_of* relationships. We have undertaken two approaches to mark such relationships for unambigous placement of entities when merging relations from one or many sentences.

3.2.4 Links between entities

1. Sometimes the same entity appears in multiple forms in the sentence. We annotate them with a "synonym" link. Sometimes, one of the forms is just an acronym for another form, in which case we use "abbreviation" link.

2. Biologically nested entities are linked with a *part_of* link. For example, protein-domains and protein-regions are part of proteins, while protein subunits are part of complexes. These links correspond to the substrate chemical structure ontology presented in Genia dataset (Thompson et al., 2017).

3.2.5 Grounding

Entities of types gene, protein, protein-family and chemical have been cross-referenced with external database identifiers. The aim of grounding is to introduce unique naming/identification of entities. This is particularly useful for unambiguous identification of entities in the process of merging relations derived from different sentences into a single network.

Notably, as a side effect, the process of grounding increased the quality of entity annotation for the specified entity types.

3.2.6 Ambiguities

Entity annotation is not a straightforward task, as entities usually appear in a variety of grammatical and biological forms. Therefore, we have developed the following guidelines for standardized annotations. Formation of these guidelines was a result of iterative annotations followed by resolution of inter-annotator conflicts.

1. *Entity modifications*

   Sometimes the text contains an entity which is a mutated form of another entity, or it is an entity in an unusual state. In these cases we tag the entity with "mutant" and/or "state" labels (Appendix A.1, example 1).

2. *Spanned and nested entities*

   If an entity contains multiple tokens, those may be separated by other words in the text, or may themselves contain nested simpler entities. In cases when the same token is shared between multiple complex entities, we annotate the shared tokens only as part of the first entity (Appendix A.1, example 2). A better solution to these cases would be to annotate the shared tokens in all the entities that they are part of and use a text-span notation to mark those cases. However, considering the small number of such cases, we didn't find this worthwhile. Sometimes a complex entity name contains a name of another entity. We annotate both, and both can appear in interactions. In extreme cases, the second entity can be a single digit. In contrast to our approach, entity recognizer systems that do not support nested entities are not be able to find these cases. In evaluation, we have a separate score that reports performance on the nested entities (Appendix A.1, examples 3-5).

3. *A/B syntax*

   In many cases A/B means a complex of the proteins A and B. In other cases it refers to separate proteins A and B, and the interaction with A/B means interactions with both of them. In both cases, we annotate A and B as individual entities. In case of complexes, we also annotate A/B as a complex. If A/B is involved in an interaction with a protein C, we annotate an interaction between A/B and C only if A/B is a complex. If A/B is not a complex, we annotate two interactions between A and C, and B and C. (Appendix A.1, example 6)

4. *Hidden entity names and implicit coreferences*

   Sometimes the sentence is about an entity which is not explicitly mentioned, but there are words that refer to it. We do not annotate these words as entities and do not annotate corresponding interactions (Appendix A.1, example 7).

3.3 Interactions

We annotate binding interactions between several types of entities.

3.3.1 Interaction types

We use three labels: 1 if the interaction exists, 0 for speculations (if the sentence does not conclude whether the interaction exists or not), and -1 for negations (if the sentence concludes that there is no binding interaction between the entities).

We conclude that an interaction exists (1) if we find explicit triggers describing direct physical interactions, such as *A binds/ associates with/ interacts with /recuits /phosphorylates B*, and their grammatical varieties.

Speculative interactions (Appendix A.2, examples 1-2) arise either due to lack of experimental evidence or due to the sentence not reaching the conlusion yet. We mark such cases with a "hypothesis" label. Other cases may be sentences that are actually titles of the sections or even the papers. In practice, title of the paper might be extracted both from the title section of the paper and from the reference sections of other papers. We tag the sentences extracted from paper, section or figure titles by "title" label (Appendix A.2, examples 3-4).

3.3.2 Ambiguities

1. *Entity polymorphisms*

   When an entity participating in an interaction appears in multiple forms in the sentence (e.g. plural forms, synonyms, etc.), we annotate the one which is the most obvious from the sentence. In evaluation, we do not penalize the predictions with another form of the same entity (Appendix A.2, example 5).

2. *Static interactions: protein complexes and domains*

   Static or implicit interactions refer to cases where an interaction is inferred from the context, but is not mentioned with any explicit trigger.

When the sentence contains a complex of two or more proteins, and the components of the complex are present in the sentence, we annotate a binding interaction between them and tag it with a “complex” label. In rare cases, the same sentence contains another explicit mention of the interaction between two proteins. In this cases we do not tag the interaction with “complex” label (Appendix A.2, examples 6-7). In evaluation, we additionally report the performance on such implicit binding interactions inside complexes.

Sometimes we annotate a (positive) binding interaction between entities A and B, where B is a region (*part_of*) of another entity C. The most common scenario is when B is a protein domain and A and C are proteins. In this case, we annotate another interaction between A and C and tag it with an “implicit” label. The full list of entity types that can get involved in similar implicit interactions is presented in Appendix A.4. We have automatically verified that all such implicit interactions are annotated (Appendix A.2, examples 8-9).

3. *Self interactions*

   There are cases when an entity binds to itself, especially when the entity is a protein-family and the binding can refer to different members of the same family (Appendix A.2, example 10).

   In rare cases, the sentence talks about homodimers or oligomerization, which implies that there is a protein which binds to itself. We tag these cases with an “implicit” label (Appendix A.2, example 11-12).

4. *Interactions with implicit entities*

   Sometimes the sentences contain interactions with entities without naming them. We exclude these interactions from the dataset (A.2, example 13).

3.4 Dataset statistics

The lengths of sentences vary from 3 to 138. The median length is 29, the mean is around 30. 95% of all sentences have less than 50. The average number of entity clusters per sentence is 3.92, while the average number of entity mentions per sentence is 4.91. On average, there are 1.61 interaction per sentence.

We used Cytoscape (Shannon et al., 2003) to construct a graph based on positive interactions annotated from our dataset. It has 2248 nodes (entities) and 3235 edges (interactions) (see Figure 2 in Appendix A.5). The graph had a large connected component, containing 65% (1475) of nodes and 81% (2635) of edges. Many interactions were annotated multiple times, with 67% (2177) of unique interactions, and up to 11 duplications per entity pair. The graph showed small-world properties, with average shortest path between any pairs of nodes being 5, and with very few hub nodes. Degrees range from 1 to 83 with median 1.

3.5 Comparison with other datasets

Table 1 compares BioRelEx 1.0 with the popular related datasets. The original version of AIMed has similar number of sentences to BioRelEx, but the number of annotated relations is significantly lower due to different annotation guidelines and choice of sentences. BioInfer contains fewer sentences with a lot more detailed annotations, which is not suitable for the current machine learning techniques, hence most of the models designed for BioInfer simply ignore the details of annotations. Both datasets do not have corresponding well-defined benchmarks. The five datasets in a unified format from (Pyysalo et al., 2008a) suit better for machine learning research, but they are limited to relation classification tasks.

The dataset for BioCreative VI Precision Medicine Track has 6.5 times more sentences than BioRelEx 1.0, but has two times less relations, as it is focused on a more rare kind of interactions.

GENIA corpus is the closest in spirit to ours. It has more detailed annotations and covers more relation types. As a result, the density of binding interactions in GENIA is much lower (only 2448 binding interactions in 9372 sentences). Also, there is a slight difference in the goals of GENIA and BioRelEx. GENIA is best suited for functional annotation and biomedical search optimization. We however, had a different aim in mind - to retrieve interactions in a way to make them useful for interaction network generation. This difference affected the way we have designed the annotation guidelines, as described in the previous subsections. Because of these differences we did not use the ontologies developed in GENIA.

In contrast to all mentioned datasets, BioRelEx includes grounding information for most of the labeled entities.

4 Benchmark

We propose a relation extraction benchmark on top of our dataset. The task is to take the raw text input and produce clusters of entity mentions along with binding interactions between the clusters. We define two main evaluation metrics, one for entity recognition and one for relation extraction. In addition to these, we define several other evaluation metrics that can be helpful in error analysis.

The main evaluation metrics are:

- Entity recognition performance in terms of micro-averaged precision, recall and F-score. In this metric we count each occurence of an entity as a separate item, and measure if the system could find all mentions in the sentence.

- Relation extraction performance in terms of micro-averaged precision, recall and F-score. Relation extraction is measured between entity clusters. Each cluster can be represented by multiple entity names in the sentence. We consider a relation between two entity clusters correctly detected, if the system predicts a relation between all pairs of entity names from the two clusters.

Two common problems of experimental setups used in relation extraction literature, as described in (Pyysalo et al., 2008b), are the inconsistent training/dev/test splits and hyperparameter tuning on the test set. To prevent these issues, we enforce a precise evaluation procedure. Following (Luan et al., 2018), we randomly split the dataset into training/dev/test sets with 70%/10%/20% ratio. The training, dev and test sets contain 1405, 201 and 404 sentences, respectively. Training and dev parts are publicly available as JSON files. We will set up a publicly available evaluation server to ensure having a truly blind test set. Additionally, we have released the evalution script used in the server[1]. We encourage everyone to use the dev set for model selection only.

4.1 Error analysis

To help with error analysis, we propose few more evaluation metrics.

Entity names: Each entity name can be mentioned multiple times in the sentence. If a model finds only one of the mentions, it is considered as a match for this score. This metric helps to verify the consistency of entity recognition in different parts of the sentence.

Flat entities: Many relation extraction systems do not support recognition of nested entities. This score acts as if there are no flat entities. More precisely, we do two modifications before calculating precision and recall:

1. If an entity mention was found by a system, we remove all entity mentions that intersect with that one from the prediction and ground truth.

2. For the remaining entity mentions we keep only the ones which do not contain another mention (e.g., only shortest mentions).

Entity coreferences: Sometimes, several entity names refer to the same actual entity. For each sentence we construct a graph, where entity names are the vertices, and two vertices are joined with an edge if they refer to the same underlying entity (are synonyms or abbreviations). This graph consists of one or more connected components, where each component is a clique and refers to a single unique entity. We measure precision, recall and f-score of the edges of the abovementioned graph. This metric helps to measure the impact of synonym or abbreviation detection.

Relation extraction (any): This metric measures relation extraction in a weaker form. We consider a relation between two entity clusters correctly detected, if the system predicts a relation between any pair of entity names from the two clusters.

Relation extraction (positive): Annotated relations have one of the three labels: 1 if the sentence confirms there is an interaction, -1 if the sentence confirms there is no interaction, and 0 if the sentence is inconclusive. We report scores that do not penalize if relations with labels 0 or -1 are not detected.

Relation extraction (non-implicit): Some of the interactions are marked as "implicit" by the annotators. These are the interactions which can be

[1] The dataset files along with the description of the JSON structure and the evaluation scripts are available at `https://github.com/YerevaNN/BioRelEx/`

	Task	Split	Relation Types	Sentences	Entities	Relations
AIMed (Bunescu et al., 2005)	Relation extraction	No	No	1978	4141	816
BioInfer (Pyysalo et al., 2007)	Relation extraction	No	Ontology	1100	6349	2662
AIMed* (Bunescu et al., 2005)	Classification	Yes	No	1955	4301	978
BioInfer* (Pyysalo et al., 2007)	Classification	Yes	No	1100	6349	2662
HPRD50* (Fundel et al., 2006)	Classification	Yes	No	145	406	160
IEPA* (Ding et al., 2001)	Classification	Yes	No	486	1118	340
LLL* (Nédellec, 2005)	Classification	Yes	No	77	239	162
BioC V BEL (Rinaldi et al., 2016)	BEL extraction	Yes	Yes	6353	N/A	11066
BioC VI PM (Dogan et al., 2017)	Relation Extraction	Yes	No	12751	10325	1629
BioNLP GE (Kim et al., 2003)	Classification+Coref	Yes	Ontology	9372	93293	36114
BioRelEx 1.0	Relation Extraction	Yes	Only binding	2010	9871	3235

Table 1: Comparison of BioRelEx 1.0 with the most popular protein-protein interaction datasets. The ones mentioned by asterisk are the unified versions from (Pyysalo et al., 2008a)

hard to detect, as they require relatively complex reasoning. We report scores that do not penalize if an implicit interaction is not detected.

All our evaluation scripts use test set bootstrapping to compute confidence intervals for the scores and to test whether the difference between two models is significant.

5 Experiments

5.1 Baselines

We provide several baselines for the benchmark described in the previous section. First, we report several trivial baselines with gold standard entities, as well as using an off-the-shelf named entity recognizer. Next, we evaluate REACH, an end-to-end biological relation extraction system, which does not require re-training. Finally, we train SciIE, an end-to-end neural network which is known to produce state-of-the-art results on similar tasks.

5.1.1 Trivial baselines

Following (Pyysalo et al., 2008a), we report scores produced by co-occurence baselines. First, we take all gold entities from the dataset and assume that there are binding interactions between all of them. This baseline gives a perfect recall and is called “Co-occur (gold)”. Then, we pass the sentences to a biomedical named entity recognition system `SciSpacy` (Neumann et al., 2019) (trained on JNLPBA corpus) and assume that there are binding interactions between all pairs. This baseline is called “Co-occur (SciSpacy)”.

5.1.2 REACH

REACH (Valenzuela-Escárcega et al., 2018) is a rule-based relation extraction system The authors host a web-based service for extracting relations from biomedical texts. We did not train or tune the system. The technical details on how we evaluated REACH system on our dataset is presented in Appendix A.6.

5.1.3 SciIE model

SciIE (Luan et al., 2018) is a complex multitask neural architecture developed by University of Washington for relation extraction from computer science paper abstracts. The model produces candidate spans of tokens, and then attempts to jointly predict entities, coreferences and relations between entities based on the spans. SciIE supports multi-word and nested entities. The technical details about adapting our data for SciIE architecture are available in Appendix A.7.

5.2 Results

The results of the four baselines on the test set of BioRelEx 1.0 are presented in Table 2. If the entity names are known, getting 35% F-score for relation extraction is trivial. Recall for relation extraction of the co-occurrence baseline is less than 100% because of the self interactions in the dataset. On the other hand, entity recognition is not easy. SciSpacy’s named entity recognizer trained on the famous JNLPBA dataset (derived from GENIA corpus) gets 67% precision and less than 53% recall. Part of the low recall is because SciSpacy’s NER cannot produce nested entities. The co-occurrence baseline with these entities gets less than 20% F-score for relation extraction.

SciIE model has a large number of hyperparameters. We kept the values mentioned in the official repository for SciERC dataset with one exception: we have changed `max_arg_width` to 5, as

		Entity Recognition	Relation Extraction	Co-occur (SciSpacy)	Co-occur (Gold)	REACH	SciIE
Co-occur (SciSpacy)	P	67.3 ± 1.4 (64.6 − 69.8)	12.6 ± 1.3 (10.3 − 15.2)		0.0%	0.2%	0.0%
	R	52.6 ± 1.5 (49.8 − 55.5)	45.1 ± 3.7 (38.5 − 52.3)				
	F_1	59.0 ± 1.3 (56.4 − 61.6)	19.6 ± 1.9 (16.3 − 23.5)				
Co-occur (Gold)	P	100.0 ± 0.0 (100 − 100)	21.5 ± 1.3 (19.2 − 24.2)	100.0%		64.8%	0.0%
	R	100.0 ± 0.0 (100 − 100)	99.2 ± 0.5 (98.1 − 99.9)				
	F_1	100.0 ± 0.0 (100 − 100)	35.3 ± 1.8 (32.2 − 38.9)				
REACH	P	70.6 ± 1.4 (68.1 − 73.1)	63.2 ± 3.9 (55.6 − 70.7)	99.8%	35.2%		0.0%
	R	65.9 ± 1.3 (63.4 − 68.3)	23.2 ± 2.3 (19.1 − 27.6)				
	F_1	68.2 ± 1.1 (65.9 − 70.3)	33.9 ± 2.8 (28.6 − 39.2)				
SciIE	P	87.7 ± 1.0 (85.8 − 89.6)	53.2 ± 2.3 (48.9 − 57.9)	100.0%	100.0%	100.0%	
	R	63.3 ± 1.6 (60.2 − 66.3)	47.4 ± 3.1 (41.1 − 53.1)				
	F_1	73.5 ± 1.3 (71.0 − 75.8)	50.1 ± 2.3 (45.5 − 54.3)				

Table 2: Results of the four baselines on the test set of BioRelEx 1.0. We report precision (P), recall (R) and F-score (F_1) for entity recognition and relation extraction. Every metric is calculated $n = 1000$ times by bootstrapping on the test set. The table shows mean, standard deviation and 95% confidence interval of 1000 runs. The right part of the table shows how often one baseline beats the other ones in 1000 evaluations according to F-score of relation extraction. We consider the difference between two models to be significant if one performs better than the other in 95% of cases.

there are very few entities with more than five tokens. We did several experiments with different weights for the NER and coreference branches of the model and picked the combination which performed best on the dev set of our dataset.

SciIE model significantly outperforms REACH system on the F-score of relation extraction: 50.1% vs 33.9%. On the other hand, REACH has a better precision for relation extraction. The difference between REACH and co-occurrence baseline with gold entities is not significant.

5.3 Error analysis

To measure the impact of nested entities on entity prediction performance we calculate **Flat entities** metric and compare it with the main entity recognition metrics. Recall jumps from 65.8% to 71.2% for REACH and from 63.3% to 68.9% for SciIE.

Our error analysis tools measure coreference detection performance. Both REACH and SciIE baselines do not output coreferences. SciIE is capable of producing coreference clusters, but the best performance on the dev set.

The relaxed versions of relation extraction evaluation do not change the results significantly. In particular, **Relation extraction (any)** metric gives 35.5% (vs. 33.9%) for REACH and 51.0% (vs. 50.1%) for SciIE.

To understand the impact of sentence lengths on the performance of the models we calculate our main metrics on the top and bottom halves of the list of sentences from dev set sorted by length. For REACH, F-score on longer sentences is worse by 1.2 and 0.8 percentage points for entity recognition and relation extraction, respectively. For SciIE, the differences are much larger, 7.4 and 9.9 percentage points respectively.

5.4 Qualitative analysis

To understand how the SciIE baseline model performs in real-world settings, we did the following experiment. We took a figure from a paper that describes MAPK-ERK signaling pathway. Figure 1a shows the schematic representation of the pathway, as described in the paper (Dantonio et al., 2018). The caption of the figure in the original paper reads: “In regular conditions, ligands such as growth factors or mitogens bind to the RTK, which is activated by autophosphorylation. Phosphotyrosine residues recruit adaptor protein Grb2 and Sos, promoting Ras:GTP association. Activated by GAPs such as NF1, Ras hydrolyzes GTP and activates Raf, the first effector kinase in the MAPK pathway. Raf then phosphorylates MEK, which in turn phosphorylates ERK. p-ERK activates cytoplasmic and nuclear substrates”.

Figure 1b shows the network extracted by our SciIE model from the original caption with no modifications. The original scheme is depicted as an underlay with light gray shades. The true positive entities and interactions are highlighted in red.

Our dataset is biased towards sentences with the verb “bind”. To see how it affects the performance of our model, we have replaced three triggers in

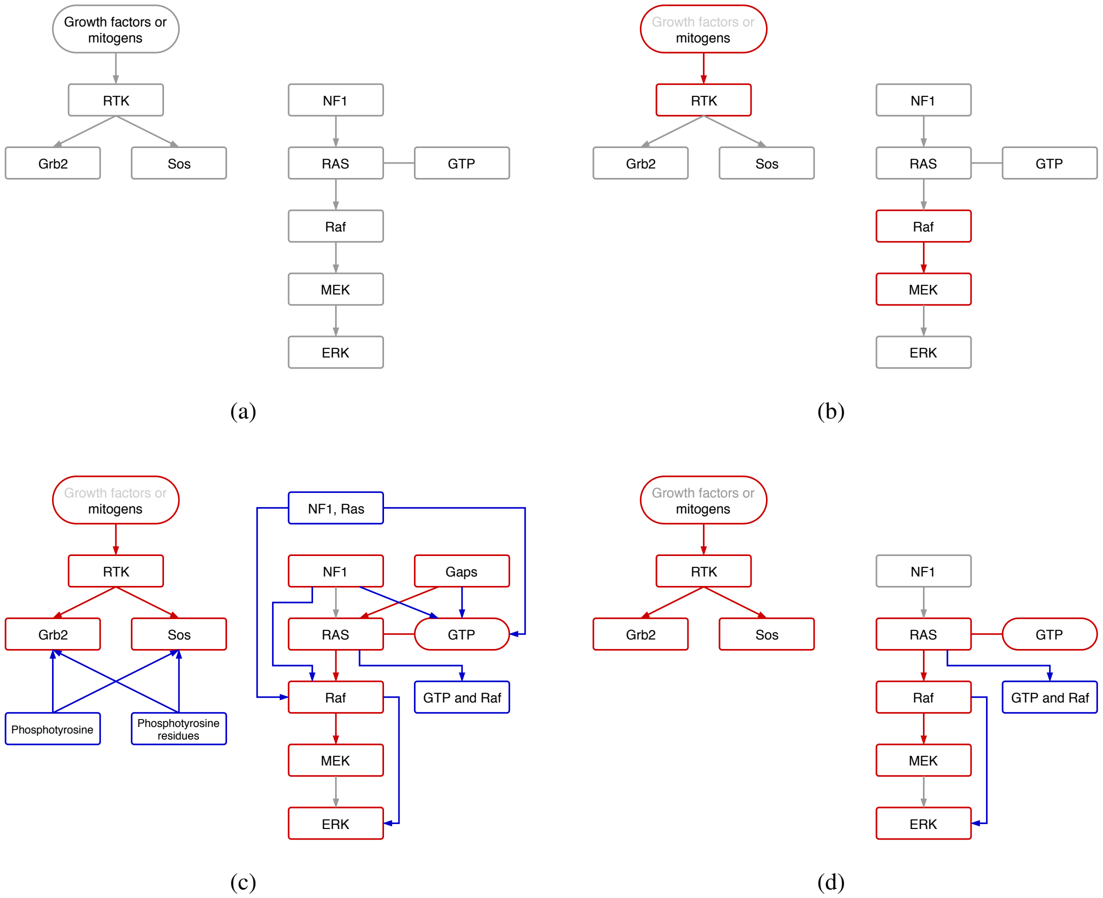

Figure 1: A network extracted by SciIE model. Refer to Section 5.4 for the details.

the original caption with "binding". The resulting network produced by SciIE is presented in Figure 1c. True positives are highlighted with red, while false positives - with blue. Note that many false entities, such as "NF1, Ras" are extracted in this case.

Finally, we removed the sentence containing the misleading "NF1", and replaced the "which in turn" coreference with "MEK". Additionally, the "phospohorylated residues" were replaced by "phosphorylated RTK" to hint the model that these residues belong to RTK. The network produced by SciIE on this version is shown in Figure 1d. The full captions used in these experiments are shown in Appendix A.8.

The results demonstrate that our SciIE baseline works much better when the interactions are expressed with the verb "bind". Additionally, we see that the lack of coreference resolution between sentences severely limits the applications of this model.

6 Conclusion

In this paper we have introduced BioRelEx 1.0, a manually annotated corpus for interaction extraction from biomedical literature. We have developed detailed guidelines for annotating binding interactions between various biological entities. The dataset is publicly available at https://github.com/YerevaNN/BioRelEx/. Based on the dataset we have designed a benchmark and evaluated several baselines on it. Finally, we have demonstrated the quality of a neural relation extraction model trained on the dataset in a real-world setting. We hope this benchmark will help to develop more accurate methods for relation extraction from unstructured text.

7 Acknowledgments

We would like to thank Sahil Garg and Martin Mirakyan for their help in the project. We would like to thank NVIDIA for donating Titan V GPUs used in the experiments.

References

Mahtab Ahmed, Jumayel Islam, Muhammad Rifayat Samee, and Robert E Mercer. 2019. Identifying protein-protein interaction using tree lstm and structured attention. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 224–231. IEEE.

Andrey Rzhetsky et. al. 2019. *in preparation*.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. Semeval 2017 task 10: Scienceie - extracting keyphrases and relations from scientific publications. *CoRR*, abs/1704.02853.

Razvan Bunescu, Ruifang Ge, Rohit J Kate, Edward M Marcotte, Raymond J Mooney, Arun K Ramani, and Yuk Wah Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial intelligence in medicine*, 33(2):139–155.

Paola M. Dantonio, Marianne O. Klein, Maria Renata V.B. Freire, Camila N. Araujo, Ana Carolina Chiacetti, and Ricardo G. Correa. 2018. Exploring major signaling cascades in melanomagenesis: a rationale route for targetted skin cancer therapy. *Bioscience Reports*, 38(5).

Louise Deléger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferré, Philippe Bessières, and Claire Nédellec. 2016. Overview of the bacteria biotope task at bionlp shared task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pages 12–22, Berlin, Germany. Association for Computational Linguistics.

Jing Ding, Daniel Berleant, Dan Nettleton, and Eve Wurtele. 2001. Mining medline: abstracts, sentences, or phrases? In *Biocomputing 2002*, pages 326–337. World Scientific.

Rezarta Islamaj Dogan, Andrew Chatr-aryamontri, Sun Kim, Chih-Hsuan Wei, Yifan Peng, Donald Comeau, and Zhiyong Lu. 2017. Biocreative vi precision medicine track: creating a training corpus for mining protein-protein interactions affected by mutations. In *BioNLP 2017*, pages 171–175.

Katrin Fundel, Robert Küffner, and Ralf Zimmer. 2006. Relexrelation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

Jeremy Getman, Joe Ellis, Stephanie Strassel, Zhiyi Song, and Jennifer Tracey. 2018. Laying the groundwork for knowledge base population: Nine years of linguistic resources for tac kbp. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885–892.

Yu Lun Hsieh, Yung-Chun Chang, Nai Wen Chang, and Wen Lian Hsu. 2017. Identifying protein-protein interactions in biomedical literature using recurrent neural networks with long short-term memory. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 240–245. Asian Federation of Natural Language Processing.

J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpusa semantically annotated corpus for biotextmining. *Bioinformatics*, 19:i180–i182.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9, Boulder, Colorado. Association for Computational Linguistics.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 1–6, Portland, Oregon, USA. Association for Computational Linguistics.

Martin Krallinger, Florian Leitner, Carlos Rodriguez-Penagos, and Alfonso Valencia. 2008. Overview of the protein-protein interaction annotation extraction task of biocreative ii. *Genome biology*, 9(2):S4.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232. Association for Computational Linguistics.

Claire Nédellec. 2005. Learning language in logic-genic interaction extraction challenge. In *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)*, volume 7, pages 1–7. Citeseer.

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of bionlp shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing.

Rose Oughtred, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, Lara ODonnell, Genie Leung, Rochelle McAdam, Frederick Zhang, Sonam Dolma, Andrew Willems, Jasmin Coulombe-Huntington, Andrew Chatr-aryamontri, Kara Dolinski, and Mike Tyers. 2018. The BioGRID interaction database: 2019 update. *Nucleic Acids Research*, 47(D1):D529–D541.

Sampo Pyysalo, Antti Airola, Juho Heimonen, Jari Björne, Filip Ginter, and Tapio Salakoski. 2008a. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(3):S6.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC bioinformatics*, 8(1):50.

Sampo Pyysalo, Rune Sætre, Junichi Tsujii, and Tapio Salakoski. 2008b. Why biomedical relation extraction results are incomparable and what to do about it. In *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008). Turku*, pages 149–152. Citeseer.

Fabio Rinaldi, Tilia Renate Ellendorff, Sumit Madan, Simon Clematide, Adrian Van der Lek, Theo Mevissen, and Juliane Fluck. 2016. Biocreative v track 4: a shared task for the extraction of causal network information using the biological expression language. *Database*, 2016.

Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504.

Paul Thompson, Sophia Ananiadou, and Junichi Tsujii. 2017. The genia corpus: Annotation levels and applications. In *Handbook of Linguistic Annotation*, pages 1395–1432. Springer.

Marco A Valenzuela-Escárcega, Özgün Babur, Gus Hahn-Powell, Dane Bell, Thomas Hicks, Enrique Noriega-Atala, Xia Wang, Mihai Surdeanu, Emek Demir, and Clayton T Morrison. 2018. Large-scale automated machine reading discovers new cancer driving mechanisms. *Database: The Journal of Biological Databases and Curation.*

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45.

A Appendices

A.1 Examples of entity annotation ambiguities

1. (`1.0.train.104`) “The inability of tyrosine-phosphorylated SLP-76 to interact with nck(SH2*)”. We annotate “nck” as a protein and “nck(SH2*)” as a protein with label “mutant”.

2. (`1.0.train.45`) “... equal amounts of REGs α and β bound to the proteasome ...”. We annotate “REGs α” as a protein and link it to the implicit “REG α”, and annotate “β” as a protein and link it to the implicit “REG β”.

3. (`1.0.dev.118`) “NF-Y binds the HSP70 promoter in vivo.”. We annotate three entities in this sentence: “NF-Y” is a protein, “HSP70” is a gene, and “HSP70 promoter” is a DNA. There is a binding interaction between “NF-Y” and “HSP70 promoter”, but not with “HSP70”.

4. (`1.0.train.964`) “...the binding of cortactin to the Arp2/3 complex....”. Here, “Arp2/3” is annotated as a complex, “Arp2” is a protein, while “3” is annotated as a protein and is linked to an implicit entity “Arp3”.

5. (`1.0.train.430`) “A18 hnRNP Binds Specifically to RPA2 and Thioredoxin 3’-UTRs”. Here, “RPA2 3’-UTR” is a region of “RPA2” RNA. But it is not a continuous span of characters, so we are forced to annotate only “RPA2”. As a result, the same sequence of characters “RPA” is annotated both as an RNA and as an RNA-region.

6. (`1.0.train.121`) “JNK/SAPK Binds and Phosphorylates a MEKK1 Fragment In Vitro”. Here JNK and SAPK are separate entities. We annotate binding interaction between “JNK” and “MEKK1” and between “SAPK” and “MEKK”.

7. (`1.0.train.50`) “... Apaf-1 binds cytochrome c and dATP, and this complex recruits caspase-9 ...”. “This complex” refers to an implicit complex with three entities. We do not annotate the complex and its interactions.

A.2 Examples of interaction annotation ambiguities

1. (`1.0.train.540`) “We also attempted to examine the actin-binding ability of partially phosphorylated F-rad.” This sentence motivates the performed experiment, but does not talk about the outcome.

Space, full stop	“S. cerevisiae” (cell), “S-1.MgADP.Pi” (protein-domain)
Question mark	No examples
Comma, colon, semicolon	“PI(4,5)P2” (chemical), “f:TFIID” (fusion-protein)
Round brackets	“NAD(H)” (chemical), “HMG-1(A-B)” (protein region)
Square brackets	“DB[a,l]PDE” (chemical), “[3H]LY341495” (drug)
Hyphen-like symbols	“IGF-II promoter” (DNA), “hTcf-4-(180)” (protein-region)
Apostrophe	“3’UTR” (RNA), “3’dE5” (chemical)
Asterisk	“Rh*” (protein), “C2A* mutant” (protein-domain)
Plus	“Ca2+” (chemical), “Na+,K+-ATPase” (protein-complex)
Dot-like symbols	“DBAD” (protein-region), “actinφ” (protein-family)

Table 3: All entities in the dataset are surrounded by any of the symbols described in the first column. On the other hand, most of these symbols can appear inside entity names. The second column of the table shows examples of entities which contain these symbols.

2. (`1.0.train.755`) “We expect that in the intact BAF complex, the actin monomer is bound to Brg1 at both of these sites.” This sentence does not confirm the existence of a binding interaction.

3. (`1.0.train.1397`) “Binding of Hairy derivatives to Gro in vitro.”. This is a title that uses an indefinite verb, and the contents of the following paragraphs might imply both existence and non-existence of the binding interaction. We annotate the binding interaction between “Hairy derivatives” and “Gro” with label 0.

4. (`1.0.train.1234`) “Phosphorylation of L1 Y1176 inhibits L1 binding to AP-2.” This is a subsection title, but it clearly implies that “L1” binds to “AP-2” (which is inhibited by phosphorylation), so we annotate this interaction with label 1.

5. (`1.0.train.1154`) “... the ORC-Cdc6p complex (and perhaps other proteins) recruits the six minichromosome maintenance (MCM) proteins ...”. Here “minichromosome maintenance” and “MCM” refer to the same protein family and are annotated as synonyms. We annotate binding interaction between “MCM” and “ORC-Cdc6p”, and the evaluation script does not penalize the model if it predicts an interaction between “minichromosome maintenance” and “ORC-Cdc6p”.

6. (`1.0.train.785`) “... TR/RXR binds to the TRE ...”. Here we annotate a binding interaction between ”TR” and ”RXR” and tag it as “complex”.

7. (`1.0.train.1154`) “... Cdc6p most likely binds to ORC and then the ORC-Cdc6p complex ...”. Here the binding interaction between “ORC” and “Cdc6p” can be inferred explicitly from the first part of the sentence and implicitly from the name of the complex. In these cases we do not tag the interaction with “complex” label.

8. (`1.0.train.630`) “hTcf-4-(180) interacts directly with the Armadillo repeats of β-catenin”. Here “hTcf-4-(180)” is annotated as a domain of “hTcf-4” protein, and “Armadillo repeats” is annotated as a region of “β-catenin” protein. We annotate the interaction between “hTcf-4-(180)” and “Armadillo repeats”. Additionally, we annotate three other interactions: “hTcf-4-(180)” and “β-catenin”, “hTcf-4” and “Armadillo repeats”, “hTcf-4” and “β-catenin”, and tag them with an “implicit” label.

9. (`1.0.train.758`) “Synaptotagmin binds β-SNAP, but not α-SNAP...”. Here “Synaptotagmin” and “SNAP” are annotated as proteins, while “α-SNAP” and “β-SNAP” are annotated as isoforms of “SNAP”. We annotate a negative binding interaction between “α-SNAP” and “Synaptotagmin”, but it does not imply that “Synaptotagmin” does not bind “SNAP”. This shows that the implicit “transfer” of an interaction does not hold if the interaction is negative.

10. (`1.0.test.171`) “Myozenin binds to both

α-actinin-2 and -3 but not to itself, whereas α-actinin-2 and -3 both bind to myozenin as well as to themselves." In this sentence we annotate a negative interaction between "Myozenin" and "Myozenin", and another positive interaction between "α-actinin-2" and "α-actinin-2".

11. (`1.0.dev.95`) "... thereby inhibiting the binding of c-Jun homodimer to TRE." Here c-Jun homodimer implies that there is a binding interaction between "c-Jun" proteins.

12. (`1.0.train.69`) "... Shs1 can bind to Gin4 and induce Gin4 oligomerization ..." Here oligomerization implies a binding interaction between "Gin4" and "Gin4".

13. (`1.0.train.783`) "Binding of IL-1 and TNF-alpha to their receptors activates several signaling pathways, including the NFkappaB and AP-1 pathways.". We do not annotate any binding interactions in this sentence, as "IL-1 receptor" is not an explicitly mentioned entity.

A.3 Tokenization rules

Table 3 describes the tokenization rules used in BioRelEx 1.0.

A.4 Entity types

Table 5 lists all entity types with descriptions used in BioRelEx 1.0 and some useful statistics[2].

Table 4 lists the pairs of entity types that are in *part_of* relationship for which we automatically add interactions to the dataset.

A.5 BioRelEx 1.0 graph

We have constructed a graph that represents the whole annotated dataset (Fig. 2) using Cytoscape tool (Shannon et al., 2003). We use grounding information to match entities from different sentences. If grounding information is not available, we fall back to entity names.

A.6 REACH baseline

We use two API calls to get information from REACH system[3]:

Child	Parent
protein-domain	protein
protein-region	protein
protein-state	protein
protein-isoform	protein

Table 4: If the sentence contains a positive binding interaction between entities A and B, where A is of a "child" type listed in this table, and it belongs to another entity C of a corresponding "parent" type, then we additionally annotate an implicit binding interaction between B and C.

- In `fries` mode, the server outputs information about entities. Each object corresponds to one entity mention in the text. Each mention has a text, location in the text, type of the entity and grounding information. In rare cases, the same entity name has different grounding information for different locations in the text. Our system does not support this scenario, so we keep the grounding information from the first mention.

- In `indexcard` mode, the server outputs information about interactions between entities. Entities have grounding identifiers which can be matched to the output of the `fries` mode. We only take the interactions which have `binds` type. In one case this API returned an interaction, where the second participant was a list of two entities. In these cases we take the first one only.

We group multiple mentions of the same entity name by matching the string. Then we group multiple entity names into an entity cluster (`unique_entity` object) by taking into account the grounding information (the concatenation of `namespace` and `ID` from REACH output).

REACH attempts to detect many entity types. We keep only the following entity types: `celline, family, protein, simple-chemical, site`. Including other types (e.g. `bioprocess, organ`, etc.) decreases precision of entity recongition (as these are not annotated in the dataset).

The implementation of our pipeline based on REACH is available on GitHub[4].

[2] We originally annotated DNA-motifs and DNA-regions as separate entity types, but after some analysis we have seen inconsistencies: sometimes DNA-motifs were annotated as DNA. We made a decision to merge all these entity types into a single cluster with name "DNA".

[3] http://agathon.sista.arizona.edu:8080/odinweb/api/text

[4] https://github.com/YerevaNN/Relation-extraction-pipeline/

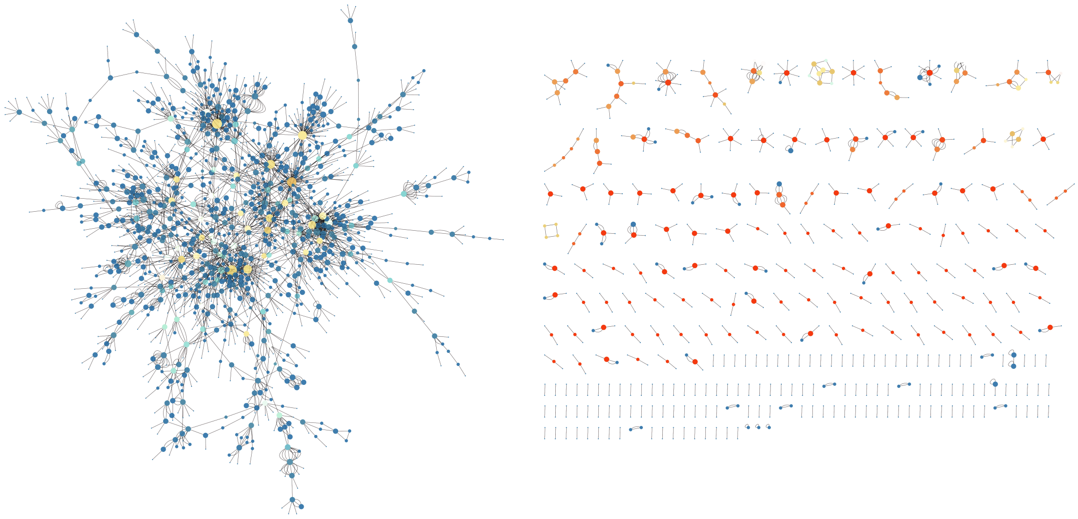

Figure 2: The network of the interactions annotated in BioRelEx 1.0

A.7 SciIE baseline

To use the SciIE model for our dataset we had to convert our data to the format the model can accept. We used a tokenzier from `scispacy` (Neumann et al., 2019), matched the tokens with our annotated entities, and added entity type, cluster (coreference) and relation information. The model supports multiple relation types. We have only one type: `bind`. Additionally, we have converted SciIE code to Python 3, and the converted version was made available on GitHub[5].

Unfortunately, the set of entities produced by the entity recognizer submodule is not syncronized with the entities that appear in the predicted coreference clusters and relations. We have developed another script to convert the output of SciIE to a JSON format that our evaluation script can handle. For entities, we used the output of SciIE entity recognizer submodule (along with the predicted entity types) and concatenated the entities that were produced by coreference and relation extraction submodules (with a label `other`). In our JSON, we specify a relation between entity clusters, although SciIE produces relations between individual entity mentions.

Our preprocessing and post-processing scripts are available on GitHub[6].

A.8 Captions for Figure 1 of Section 5.4

C In regular conditions, ligands such as growth factors or mitogens bind to the RTK, which is activated by autophosphorylation. Phosphotyrosine residues bind to adaptor protein Grb2 and Sos, promoting Ras:GTP association. Activated by GAPs such as NF1, Ras binds GTP and Raf, the first effector kinase in the MAPK pathway. Raf then phosphorylates MEK, which in turn binds ERK. p-ERK activates cytoplasmic and nuclear substrates.

D In regular conditions, ligands such as growth factors or mitogens bind to the RTK, which is activated by autophosphorylation. Phosphotyrosine residues bind to adaptor protein Grb2 and Sos, promoting Ras:GTP association. Ras binds GTP and Raf, the first effector kinase in the MAPK pathway. Raf then phosphorylates MEK, afterwards MEK binds ERK. p-ERK activates cytoplasmic and nuclear substrates.

[5] https://github.com/YerevaNN/SciERC/

[6] https://github.com/YerevaNN/Relation-extraction-pipeline/

Entity type	Statistics	Description
protein	3640 / 3777 / 82 / 147	Entities either represented with protein names; or with gene names (X) but factually standing as actual proteins in the sentence (either explicitly: X protein; or implicitly X binds the promoter)
protein-family	1086 / 1017 / 0 / 0	Entities represented with protein-family names (e.g. actin) or representing a group of protein with common properties (e.g. globular proteins; x-domain containing proteins, etc)
chemical	532 / 295 / 0 / 0	Any chemical compound other than protein or DNA or RNA, excluding experimental reagents/antibodies.
DNA	506 / 468 / 2 / 0	Any entity type that represents a region of or full DNA molecule, except for gene names. These include explicit 'DNA' mentions; DNA-regions, such as gene promoters, DNA elements; DNA sequences represented with nucleotides and DNA-motifs represented with names; chromosomes and plastids.
protein-complex	419 / 294 / 1 / 0	Protein complexes are either explicitlly mentioned with name followed by 'complex' suffix, or with name containing subunits seperated with slashes or dashes, or with names that do not contain the members, but are known to be complexes.
protein-domain	318 / 134 / 2 / 1	Domains may or may not be explicitly annotated with the suffix 'domain'. They may be specific domains of proteins present in the sentence, or general domain names without reference to the proteins they belong to.
cell	152 / 1 / 3 / 2	Explicit mentions of a cell or entities representing cell names, cell-line, bacterium, as well as viruses.
experimental-construct	141 / 60 / 0 / 0	Entities refering to artificially merged molecules, including tagged proteins, tagged RNA and DNA and chemically modified proteins/RNA/DNA.
RNA	137 / 105 / 0 / 0	All the entities representing physical RNA molecules (mRNA, tRNA, rRNA, etc.), or RNA-motifs (represented by RNA sequence or motif name) or RNA regions (represented by region names). mRNAs presented in text with corresponding gene names are also annotated as RNA.
experiment-tag	128 / 35 / 0 / 0	Chemicals or proteins experimentally added to proteins (e.g. GST tag).
reagent	128 / 43 / 0 / 0	Chemicals/biomolecules used in experimental settings (e.g. antibody)
protein-motif	122 / 43 / 0 / 0	Amino acid sequence patterns represented either by motif names or amino acid sequences, which may or may not be followed by explicit 'motif' mention.
gene	109 / 6 / 2 / 0	Entities represented with gene names.
amino-acid	69 / 2 / 0 / 0	Amino acids represented by amino acid names or explicit amino acid mentions.
protein-region	66 / 37 / 0 / 0	Protein regions are entities refering to amino-acid sequences (motif names or actual sequence representations); or regions on the protein not refering to whole domains.
assay	55 / 0 / 0 / 0	Entities refering to exprimental method names or assays or procedures.
organelle	51 / 20 / 0 / 0	Subcellular entities represented with their names (e.g. ribosome).
peptide	37 / 24 / 0 / 0	Short amino-acid polymers represented by their names, which may or may not be followed by explict 'peptide' mentions.
fusion-protein	32 / 25 / 0 / 0	Fusion-proteins
protein-isoform	32 / 33 / 0 / 0	Protein sub-types encoded by the same gene, but resulting from its differential post-processing. These entities may or may not appear in a sentence with explicit isoform mentions.
process	31 / 0 / 0 / 0	Entities refering to sequences of events at molecular, cellular or organismal levels. These may be pathway names (represented either by member gene names or target process names, with or without explicit 'pathway' mentions); process names/descriptions (e.g. autophagy); disorders and biological phenotypes.
mutation	20 / 0 / 0 / 0	Specifications of mutations in the form of nucleotide-to-nucleotide (A55G) or amino acid-to-amino acid transitions (Ala55Ser) or sequence to sequence transitions (ACGT to AGGT).
protein-RNA-complex	20 / 11 / 0 / 0	Complexes composed of proteins and RNA, mentioned either with component names or the complex alias, with or without explicit 'protein-RNA' mention.
drug	18 / 8 / 0 / 0	Drug names
organism	7 / 0 / 0 / 0	Multi-cellular organisms (i.e. excluding cells, bacteria and viruses)
disease	6 / 0 / 0 / 0	Entities representing disease names.
protein-DNA-complex	5 / 7 / 0 / 0	Complexes composed of proteins and DNA, mentioned either with component names or the complex alias, with or without explicit 'protein-DNA' mention.
brand	4 / 0 / 0 / 0	Entities representing company names or reagent/drug brands.
tissue	2 / 0 / 0 / 0	Entities representing tissues.
RNA-family	2 / 1 / 0 / 0	Entities representing groups of RNA with common properties.
gene-family	2 / 0 / 0 / 0	Entities representing sets of genes encoding for protein-families or combined by a common characteristic. Usually mentioned with name followed by 'gene family'.
fusion-gene	1 / 1 / 0 / 0	Entities representing fusion products of two genes. Usually represented by gene names separated with dashes followed (or not) by 'fusion' suffix.

Table 5: Entity types annotated in the dataset. The second column shows the number of mentions of those entities in the sentences, number of binding interactions involving those entities, number of mutated entities and number of entities that appear in a special state (e.g. phosphorylated).