Pathway Curation Support as an Information Extraction Task

Tomoko Ohta[*] **Sampo Pyysalo**[†‡] **Sophia Ananiadou**[†‡] **Jun'ichi Tsujii**[§]
[*]Department of Computer Science, University of Tokyo, Tokyo, Japan
[†]National Centre for Text Mining, Manchester, UK
[‡]School of Computer Science, University of Manchester, Manchester, UK
[§]Microsoft Research Asia, Beijing, China
okap@is.s.u-tokyo.ac.jp,sampo.pyysalo@gmail.com
sophia.ananiadou@manchester.ac.uk,jtsujii@microsoft.com

Abstract

Pathway curation, the process of synthesizing information from the biomedical scientific literature to produce formal representations of complex biological processes, is a major undertaking in molecular biology. Despite numerous studies aiming to build representations of molecular reactions from text automatically, information extraction (IE) methods have not been widely adopted in pathway curation efforts. We argue that to become more relevant to these efforts, IE methods should address the full structured representations used for pathway models and that methods should *support* manual curation efforts rather than seek to entirely automate them. In this paper, we propose a specific support task relevant to practical pathway curation efforts and evaluate its feasibility through a pathway-oriented evaluation using a recently introduced large-scale IE resource.

1 Introduction

Detailed systematic representations of the entities and reactions involved in complex biomolecular processes are critical for establishing an understanding of the molecular foundations of both physiological processes and disease (Kitano, 2002) and the production of such representations, pathway curation, is a major focus of efforts in present-day biology. Formal, computer-readable pathway representations are necessary for automatic data management and exchange, verification, database integration, and bioprocess simulation. However, pathway model curation is a very demanding task: large models incorporate information from hundreds of individual publications, and the curation and summarization of such information into pathway representations is still effectively performed by purely manual efforts.

To reduce manual effort and to maintain curation consistency, pathway curation potentially stands to benefit greatly from automatic support in the analysis of the scientific literature through advanced search, information extraction (IE) and text mining techniques (Ananiadou et al., 2010). To address these opportunities, the biomedical natural language processing community has proposed a number of systems under headings such as *automatic pathway extraction* (Park et al., 2001; Rzhetsky et al., 2004; Yao et al., 2004; Rajagopalan and Agarwal, 2005; Yuryev et al., 2006; Zhang et al., 2009). However, instead of addressing pathway models applied in curation by biologists, IE efforts have focused almost exclusively on the extraction of custom representations with different semantics, often failing to meet the requirements for detailed biological pathway models e.g. in being restricted to only pairwise entity associations (Oda et al., 2008). This may explain in part why these systems have not been widely adopted by the biological curation community.

In this paper, we aim to define a *pathway curation support* task explicitly in terms of a widely applied formal pathway representation as a step toward improving the practical applicability of IE for pathway curation. We assess the feasibility of the proposed task through a detailed analysis of a literature-scale database of information extracted by a state-or-the-art IE system from the perspective of three recently introduced pathways.

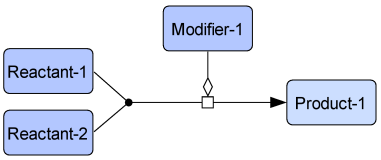

Figure 1: Illustration of a generalized pathway reaction involving two reactants, one product, and one modifier.

2 Pathways

2.1 Introduction

Pathway curation is pursued by a large number of groups in all areas of molecular biology, using a great number of different representations of varying expressibility and formality (see e.g. Kitano et al. (2005)). As most pathway representations have originally been developed independently, their compatibility and the ability to convert between representations is limited. To address the challenges of interoperability, the curation community has pursued standardization efforts, proposing a number of specific representations, resources and tools for data interchange (Hucka et al., 2003; Mi and Thomas, 2009; Demir et al., 2010; Matsuoka et al., 2010).

Here, we focus on the Systems Biology Markup Language (SBML) standard, which defines an expressive XML-based representation and has broad support in the curation community.[1] As a standard intended to support flexible data interchange, SBML does not define a fixed set of entity or reaction types, but allows different semantics to be defined via extension mechanisms. We consider specifically SBML with CellDesigner extensions (Funahashi et al., 2008), which define a rich set of entity and reaction types that are broadly compatible with other standard proposals such as BioPAX (Demir et al., 2010). CellDesigner is also applied for annotation in many large-scale pathway curation efforts (e.g. Caron et al. (2010)) and any curation support system that is compatible with the CellDesigner SBML extensions thus immediately relevant to the day-to-day efforts of biologists working on pathways.

[1]For example, the PANTHER pathway resource (`http://www.pantherdb.org/`) consists of over 165 pathway models.

Entity	Reaction
PROTEIN	CATALYSIS
COMPLEX	STATE TRANSITION
RNA	HETERODIMER ASSOCIATION
SIMPLE MOLECULE	TRANSPORT
GENE	TRANSCRIPTION

Table 1: Examples of frequent SBML/CellDesigner entity and reaction/reaction modification types in the Payao repository pathways.

2.2 Representation

The SBML representation involves two primary top-level concepts: (physical) entities[2] and reactions. Entities and reactions are assigned types, such as PROTEIN and TRANSPORT. Reactions are captured using a structured representation that associates an arbitrary number of entities, each of which is characterized as participating in the reaction in one of the general roles of *reactant*, *product*, or *modifier*. Modifier roles are further specified as having e.g. a catalyzing effect on the reaction through the assignment of the CATALYSIS type. Figure 1 illustrates a simple generalized reaction and Table 1 gives further examples of common entity and reaction/reaction modification types.

2.3 Resources

For reference pathway data, we use a selection of models from the Payao pathway repository. Payao is an SBML model tagging platform that allows a community to share and discuss models and search related literature (Matsuoka et al., 2010; Kemper et al., 2010). The Payao repository includes a wealth of models represented in SBML with CellDesigner extensions and annotated with links to the literature in the form of PMID references associated with individual entities and reactions.

For this study, we selected the TLR (Oda and Kitano, 2006), mTOR (Caron et al., 2010), and yeast cell cycle (YCC) (Kaizu et al., 2010) pathways, three large, high-quality models with extensive literature references. Each of these pathways involves over 400 reactions between over 650 unique physical entity types curated from 400 or more original publications.

[2]Termed *species* in SBML. We have chosen not to use this term to minimize possible confusion with its general-language senses.

3 Pathway curation support

Despite the substantial number of studies broadly on the topic of pathway extraction from text, there is no accepted standard task setting or evaluation data for automatic curation support. We next discuss the general requirements and challenges in establishing such a task, how these can be met, and propose a specific task setting and evaluation criteria.

3.1 The right semantics

As we argued in the introduction, for a support system to be accepted by biologists working on curation, it is critical that its semantics agree with those used in their work. This means that the task should adopt the structured representation, the reactant/product/modifier division, the state perspective (see e.g. Oda et al. (2008)) as well as the specific entity and reaction types applied in pathway models.

We emphasize that adopting the information content and types of a representation does not necessarily imply using the same representation *format*.

3.2 A well-posed problem

A major challenge for establishing any pathway curation-related task is that there are many equally correct ways of constructing a model of a specific biological pathway. A pathway model reflects the general view, scope and preferred levels of detail and granularity set by a particular curator or group of curators. The problem of creating *the* pathway for all but the smallest sets of entities or publications is inherently ill-posed; there are a practically unlimited number of specific solutions that would be acceptable to a human evaluating them after the fact. Thus to formulate a coherent, stable pathway curation support task where success can be evaluated automatically, it is necessary to frame it within the bounds of a largely fixed granularity, scope, level of detail, and related factors. As such aspects are, by definition, fixed in any existing curated resource, the most straightforward way to frame a pathway curation support task is with respect to an already (at least largely) completed pathway.

3.3 Possible tasks

From the starting point of a partially complete pathway, it is possible to formulate a number of questions where humans are likely to largely converge on a moderately-sized set of specific answers, such as

1. What entities relevant to this pathway are missing?
2. What reactions involving the included entities are missing?
3. What is the type of entity *E* / reaction *R*?

Here, we propose to focus on the task of resolving the type and participants of a "missing" reaction. We further constrain the problem by not asking the open question (2) above but rather "what reaction *involving entities* $(E_1, \ldots, E_n)$ is missing". We propose this constrained form as answers to the broader question cannot be automatically evaluated given a fixed pathway: no pathway is completely exhaustive and so cannot serve as "gold" reference data for question (2). By contrast, for the great majority of sets of entities associated in a pathway, there is only a single correct answer to the constrained question. Thus, all reactions annotated in existing pathways – tens of thousands of individual manually annotated instances – can potentially be used to create example data for training and evaluating methods for this task.

While obviously artificial in the form of determining whether an already manually annotated reaction can be automatically recreated, the proposed task has direct relevance to practical challenges in pathway curation. The entities associated with a specific biological pathway are likely to be known at the outset of an effort to model their reactions in detail; the requirements on the input are thus not expected to hinder real-world applications. A system that is able to accurately address this task would be of value to biologists working in pathway curation in at least 1) extending a partially complete pathway 2) verifying and updating existing detailed models 3) refining existing pathways employing coarse models into a detailed representation and, assuming a system returning textual evidence, 4) discovering new literature support or adding missing links to literature to pathways (Kemper et al., 2010).

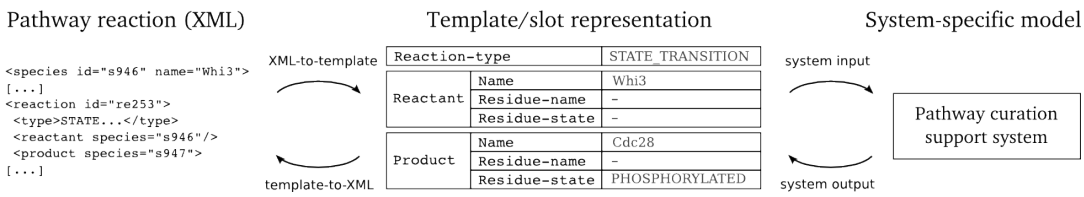

Figure 2: Proposed task setting and representations. Template slots filled by the system shown grey serif font; dashes ("-") identify empty fills. The `template-to-XML` and `XML-to-template` format conversions are deterministic and not required to address the proposed task.

3.4 Detailed task setting

We next detail the input, output and related conditions for the proposed task.

The input is a list of entity names as found in a pathway model, each of which is a simple character string.[3] In only providing entity names, additional information that may be marked in pathway models concerning entity location, activation and dimerization state, and modifications such as phosphorylation is excluded from the input: such information is frequently tied to the involved reaction and will not in practice be known unless the reaction itself is already known.[4]

The output of a system addressing the task is a representation of a pathway reaction involving the entities. As argued above, this representation should conform to the semantics of pathway models, and the most readily usable form for the output would arguably be complete, fully formatted SBML/CellDesigner XML. However, requiring such output would unnecessarily complicate the IE task, introducing additional demands on software development and reducing focus on the core problems. We propose instead to represent the reaction using a simple template/slot model represented in a a text-based format, as applied in many IE tasks (e.g. Grishman and Sundheim (1996)), defining the slots and values so that they can be deterministically mapped into full SBML/CellDesigner XML. This design assures that the simplification does not detract from the realism or applicability of the task setting.

The task setting and involved representations are illustrated in Figure 2. For the output, the system must fill a `Reaction-type` slot with a string matching the correct reaction type, and for each reactant, product and modifier a corresponding slot (e.g. `Reactant`) with values for entity name (`Name`) as well as name and state slots for any altered residues, and a slot for the type of each modification (not shown in figure). We exclude entity types from the output on the assumption that these will generally be known to users. While this specification excludes some detailed aspects of the SBML/CellDesigner representation,[5] it is sufficiently expressive to cover most reactions and can be straightforwardly extended to capture remaining aspects if needed.

3.5 Evaluation settings and criteria

We propose two alternative settings for evaluation: a *closed* setting in which the task must be addressed on the basis of a given fixed set of publications (or other texts), and an *open* setting in which any existing resource, such as the whole PubMed, can be used as reference. The closed setting assumes that it is possible to identify a subset of the literature relevant to the pathway; information that is included in many, but not all pathway models. By contrast, the open setting can be applied for any pathway, but may place considerable demands for IE systems in that it potentially requires some analysis of the entire available literature.

Systems addressing the task could be evaluated using a number of different metrics depending on perspective and the emphasis placed on different aspects of the representation. Here, we consider for each reaction separately the *core re-*

[3] As pathway models rarely normalize entity mentions (e.g. by identifying the relevant record in Uniprot), the availability of database identifiers cannot in general be assumed.

[4] For example, a reactant may have an unmodified slot that is marked as *Phosphorylated* in the corresponding product. Providing this information in input would, in effect, resolve the reaction type as phosphorylation.

[5] Specifically, entity location with respect to compartments and activity and homodimerization state are not represented.

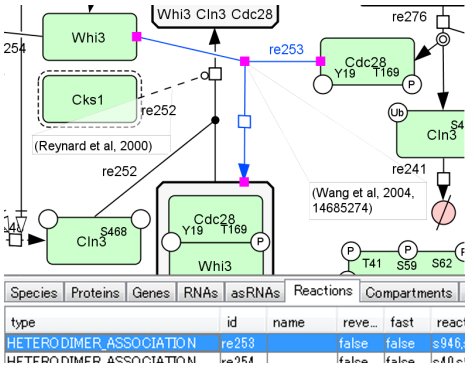

Figure 3: Example pathway reaction in CellDesigner.

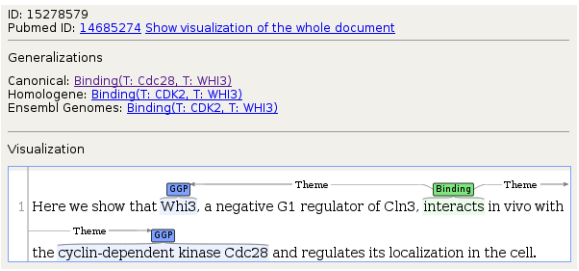

Figure 4: Example event and associated information in EVEX.

action – defined as `Reaction-type` and the `Reactant`/`Residue`/`Modifier` `Name` slots – and the *full reaction*, encompassing also modifier and residue information. Evaluation is then performed simply in terms of measuring the fraction of all reactions for which the core and full reaction information is correctly filled.

4 Experiments

We performed experiments aiming to determine the feasibility of the proposed task from two perspectives: 1) whether the abstracts of documents referenced as evidence for pathway reactions contain sufficient information to address the task, and 2) whether a state-of-the-art structured IE system can correctly extract this information.

As the proposed task setting is new and full automation of evaluation requires a number of complex software components, this feasibility study was performed entirely manually by a PhD biologist with experience in SBML/CellDesigner pathway models and domain IE.

4.1 Data

For pathway data we used the three models introduced in Section 2.3, selecting from each a random set of 100 reaction-PMID pairs for evaluation. As text data, we analysed the abstracts of the publications identified by these 300 PMID references.

As a representative of state-of-the-art domain IE, we performed evaluation against EVEX (Björne et al., 2010; Van Landeghem et al., 2011), a database resource covering the results of an automatic analysis of the approximately 20 million citations in PubMed (2009 distribution). The tools applied to create EVEX are the named entity recognizer BANNER (Leaman and Gonzalez, 2008) trained on the GENETAG corpus (Tanabe et al., 2005), the parser of Charniak and Johnson (2005) with the biomedical domain model of McClosky (2009) for analysis of sentence structure, the Turku Event Extraction System (TEES) (Björne et al., 2009; Björne et al., 2010) trained on the BioNLP Shared Task 2009 version (Kim et al., 2009) of the GENIA corpus (Ohta et al., 2002; Kim et al., 2008) for event extraction, and the method of Van Landeghem et al. (2011) for normalizing entity mentions to database identifiers. Due to space considerations we refer the interested reader to the studies presenting these various methods and resources for detailed descriptions.

Figures 3 and 4 illustrate a pathway reaction and an event structure and associated information from EVEX for a sentence stating one of the shown reactions.

4.2 Reactions in text

Table 2 summarizes the results of the initial analysis. We found that for 44% of the studied pathway reactions, the core reaction could be found stated in the abstracts of documents referenced as evidence for the reaction in whole (88% of stated reactions) or at least in part (12%). The remaining reactions were primarily missed due to the entities marked as participating in the reaction not being specifically mentioned in the abstract, or mentioned without statement of their reaction. In a small number of cases evaluation could not be performed due to missing documents[6] or due to a pathway annotation not representing a biomolec-

[6]The EVEX dataset only extends up to 2009.

Category	mTOR	TLR	YCC	Total	
Reaction stated	39	43	49	131	(44%)
No entities	27	24	24	75	(25%)
No reaction	22	25	27	74	(25%)
No document	12	0	0	12	(4%)
Non-reaction	0	8	0	8	(3%)
Total	100	100	100	300	(100%)

Table 2: Results of analysis of core pathway reactions found in the abstracts of referenced publications.

Category	mTOR	TLR	YCC	Total	
Missing modifier	5	8	5	18	(14%)
Missing residue	0	0	5	5	(4%)
Missing both	12	5	4	21	(16%)
Total	17	13	14	44	(34%)

Table 3: Results of analysis of additional aspects of 131 core reactions stated in abstracts.

ular reaction.[7] We further analysed each of the 131 cases where the core reaction was stated at least in part to determine whether any additional full reaction information, i.e. residues and modifiers, was missing from the reaction statement in the abstract. The results of this analysis are presented in Table 3.

Based on this analysis, we estimate that for a dataset including also the most recent publications, the proposed task could be resolved in approximately 50% of cases reactions for the core reaction and in approximately 33% for the full reaction based on the contents of abstracts. While encouraging for the general feasibility of the proposed task setting, this result underscores the need to perform analysis of this type to avoid the infeasible problem setting that would arise from simply assuming any annotated pathway reaction-PMID pairing represents a resolvable case for IE.

4.3 Resolving reactions using events

For the 131 reactions which were found stated in the abstracts referenced from the pathways, we next analysed the EVEX data to determine whether event extraction could support their recovery. We defined for each case a target output consisting of that subset of the reaction annotation that was found stated, identified the input names (as described in Section 3.4) and determined for each whether the defined slots could be correctly filled given only EVEX event structures *involving the input entities in the given abstract*. The results of this analysis are summarized in Tables 4 and 5.

Category	mTOR	TLR	YCC	Total	
Reaction extracted	20	29	31	80	(61%)
Out of scope	6	9	9	24	(18%)
Partial event	6	3	4	13	(10%)
No event	6	2	4	12	(9%)
No entities	1	0	1	2	(2%)
Total	39	43	49	131	(100%)

Table 4: Results of analysis of pathway reaction resolution on the basis of event structures.

Category	mTOR	TLR	YCC	Total	
Missing modifier	3	3	8	14	(34%)
Missing residue	2	1	0	3	(7%)
Missing both	1	1	3	5	(12%)
Total	6	5	11	22	(54%)

Table 5: Results of analysis of resolution of stated additional aspects of pathway reactions on the basis of event structures from which reactions could be extracted (41 cases).

Remarkably, we found that the most frequent reason why the IE outputs fail to resolve core reactions even partially is *not* due to error of the IE methods – named entity recognition (NER) or event extraction – but rather due to lack of semantic coverage. Not all pathway reaction types find correspondences in BioNLP ST'09 event types, and our evaluation indicates that "out of scope" reaction types represent a major limitation for practical, high-coverage pathway curation support. The major types for which coverage is lacking in EVEX are protein modifications, in particular Ubiquitination and Dephosphorylation, as well as the Dissociation of protein complexes. As annotated data for each of these types has been made available (Ohta et al., 2011a; Ohta et al., 2011b) since the introduction of EVEX, coverage could be extended through IE system retraining and rerunning the PubMed-scale analysis.

Of core reactions of types for which the BioNLP ST'09 event types provide coverage, a clear majority can be resolved using the EVEX data, with only 2% of cases involving a failure of the NER system to detect any relevant entity mention and 9% a corresponding complete failure of the event extraction system. While extraction per-

[7] Curators had in cases marked associations such as a protein belonging to a family as `STATE TRANSITION` "reactions", a somewhat imprecise use of the representation.

formance is notably more limited for additional aspects of reactions (Table 5), this result suggests that the basic IE technology is at least close to a level of reliability sufficient to meet the demands of real-world pathway curation support.

5 Discussion and conclusions

We have argued that to make biomedical IE technology more relevant to the production of detailed pathway models it is necessary to address pathway curation as performed in practice by biologists and to seek to support these efforts rather than supplant them. As a step toward realizing practical pathway curation support systems, we proposed a specific IE task with reference to the widely used SBML/CellDesigner pathway model. The feasibility of the proposed task setting and remaining challenges for IE technology were assessed through manual evaluation of 300 reactions from three recently introduced pathways, the literature supporting the annotation of these reactions, and automatically extracted information in a PubMed-scale resource created using state-of-the-art event extraction technology.

Analysis indicated that core reaction information is stated in the abstracts of referenced publications for approximately half of pathway reactions, suggesting that the basic task setting is feasible even without access to full texts. Study of the IE outputs further suggested that state-of-the-art event extraction technology can extract core information for over 60% of stated reactions and that remaining challenges are, surprisingly, not primarily failures of the extraction methods but rather of lack of semantic coverage. We thus urge IE method developers to consider the use of recently introduced resources addressing such omissions. Recent improvements to basic extraction technology may also be valuable in addressing lack of coverage (Kim et al., 2011).

As a first proposal of a new task and initial feasibility study, our evaluation is limited in a number of ways. Through manual analysis, we have here sidestepped some challenges relating to the mappings between text, IE outputs, the template representation, and full pathway XML. There are nontrivial engineering challenges in e.g. the required implementation of the conversion between SBML and the template representation. In allowing human interpretation of names, we have also avoided issues relating to automatic name normalization. While this task has been studied extensively and many systems are available, their performance remains somewhat limited (Lu and Wilbur, 2010). These issues are likely to restrict practical pathway curation support system performance. While our evaluation is thus an upper bound for the considered setting, this closed, single-abstract setting is arguably the most demanding possible: in a practical system, information not found in a specific abstract may still be recovered in full text or another document in a larger collection. Performing an evaluation in the open setting is key future work for further analysis of the feasibility of pathway curation support.

We hope that the resources developed as part of this effort can serve as reference data for the development and testing of systems for pathway curation support. All these resources are freely available for use in research from `http://www.geniaproject.org`

Acknowledgments

This work was patially supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan) and by the UK Biotechnology and Biological Sciences Research Council (BBSRC).

References

S. Ananiadou, S. Pyysalo, J. Tsujii, and D.B. Kell. 2010. Event extraction for systems biology by text mining the literature. *Trends Biotechnol.*, 28(7):381–390.

J. Björne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, and T. Salakoski. 2009. Extracting complex biological events with rich graph-based feature sets. In *BioNLP'09*, pages 10–18.

J. Björne, F. Ginter, S. Pyysalo, J. Tsujii, and T. Salakoski. 2010. Scaling up biomedical event extraction to the entire pubmed. In *BioNLP'10*, pages 28–36.

E. Caron, S. Ghosh, Y. Matsuoka, D. Ashton-Beaucage, M. Therrien, S. Lemieux, C. Perreault, P.P. Roux, and H. Kitano. 2010. A comprehensive map of the mTOR signaling network. *Mol Syst Biol.*, 6(1).

E. Charniak and M. Johnson. 2005. Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking. In *ACL'05*, pages 173–180.

E. Demir, M.P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D'Eustachio, C. Schaefer, J. Luciano, et al. 2010. The BioPAX community standard for pathway data sharing. *Nat Biotechnol.*, 28(9):935–942.

A. Funahashi, Y. Matsuoka, A. Jouraku, M. Morohashi, N. Kikuchi, and H. Kitano. 2008. CellDesigner 3.5: a versatile modeling tool for biochemical networks. *Proc. IEEE*, 96(8):1254–1265.

R. Grishman and B. Sundheim. 1996. Design of the MUC-6 evaluation. In *MUC-6*, pages 413–422.

M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, and H Kitano et al. 2003. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531.

K. Kaizu, S. Ghosh, Y. Matsuoka, H. Moriya, Y. Shimizu-Yoshida, and H. Kitano. 2010. A comprehensive molecular interaction map of the budding yeast cell cycle. *Mol Syst Biol.*, 6(1).

B. Kemper, T. Matsuzaki, Y. Matsuoka, Y. Tsuruoka, H. Kitano, S. Ananiadou, and J. Tsujii. 2010. PathText: a text mining integrator for biological pathway visualizations. *Bioinformatics*, 26(12):i374.

J-D. Kim, T. Ohta, and J. Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(1):10.

J-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii. 2009. Overview of BioNLP'09 Shared Task on Event Extraction. In *BioNLP'09*, pages 1–9.

J-D. Kim, S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen, and J. Tsujii. 2011. Overview of BioNLP Shared Task 2011. In *BioNLP'11*, pages 1–6, June.

H. Kitano, A. Funahashi, Y. Matsuoka, and K. Oda. 2005. Using process diagrams for the graphical representation of biological networks. *Nat Biotechnol.*, 23(8):961–966.

H. Kitano. 2002. Systems biology: a brief overview. *Science*, 295(5560):1662.

R. Leaman and G. Gonzalez. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. *PSB'08*, pages 652–663.

Z. Lu and W.J. Wilbur. 2010. Overview of BioCreAtIvE III gene normalization. In *BioCreative III*.

Y. Matsuoka, S. Ghosh, N. Kikuchi, and H. Kitano. 2010. Payao: a community platform for SBML pathway model curation. *Bioinformatics*, 26(10):1381.

D. McClosky. 2009. *Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing*. Ph.D. thesis, Brown University.

H. Mi and P. Thomas. 2009. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol. Biol*, 563:123–140.

K. Oda and H. Kitano. 2006. A comprehensive map of the toll-like receptor signaling network. *Mol Syst Biol.*, 2(1).

K. Oda, J-D. Kim, T. Ohta, D. Okanohara, T. Matsuzaki, Y. Tateisi, and J. Tsujii. 2008. New challenges for text mining: Mapping between text and manually curated pathways. *BMC Bioinformatics*, 9(Suppl 3):S5.

T Ohta, Y Tateisi, H Mima, and J Tsujii. 2002. GENIA corpus: an annotated research abstract corpus in molecular biology domain. *HLT'02*, pages 73–77.

T. Ohta, S. Pyysalo, and J. Tsujii. 2011a. From pathways to biomolecular events: Opportunities and challenges. In *BioNLP'11*, pages 105–113.

T. Ohta, S. Pyysalo, and J. Tsujii. 2011b. Overview of the Epigenetics and Post-translational Modifications (EPI) task of BioNLP Shared Task 2011. In *Proceedings BioNLP'11*.

J. C. Park, H-S. Kim, and J-J. Kim. 2001. Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. In *PSB'01*, volume 6, pages 396–407.

D. Rajagopalan and P. Agarwal. 2005. Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics*, 21(6):788.

A. Rzhetsky, I. Iossifov, T. Koike, M. Krauthammer, P. Kra, M. Morris, H. Yu, P.A. Duboué, W. Weng, W.J. Wilbur, V. Hatzivassiloglou, and C. Friedman. 2004. GeneWays: A system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J Biomed Inform.*, 37(1):43–53.

L. Tanabe, N. Xie, L.H. Thom, W. Matten, and W.J. Wilbur. 2005. GENETAG: A tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6(Suppl. 1):S3.

S. Van Landeghem, F. Ginter, Y. Van de Peer, and T. Salakoski. 2011. Evex: A pubmed-scale resource for homology-based generalization of text mining predictions. In *BioNLP'11*, pages 28–37.

D. Yao, J. Wang, Y. Lu, N. Noble, H. Sun, X. Zhu, N. Lin, D.G. Payan, M. Li, and K. Qu. 2004. Pathwayfinder: paving the way towards automatic pathway extraction. In *APBC'04*, pages 53–62.

A. Yuryev, Z. Mulyukov, E. Kotelnikova, S. Maslov, S. Egorov, A. Nikitin, N. Daraselia, and I. Mazo. 2006. Automatic pathway building in biological association networks. *BMC Bioinformatics*, 7(1):171.

L. Zhang, D. Berleant, J. Ding, T. Cao, and E. Syrkin Wurtele. 2009. PathBinder - text empirics and automatic extraction of biomolecular interactions. *BMC Bioinformatics*, 10(Suppl 11):S18.