From Pathways to Biomolecular Events: Opportunities and Challenges

Tomoko Ohta* **Sampo Pyysalo*** **Jun'ichi Tsujii†**
*Department of Computer Science, University of Tokyo, Tokyo, Japan
†Microsoft Research Asia, Beijing, China
{okap,smp}@is.s.u-tokyo.ac.jp, jtsujii@microsoft.com

Abstract

The construction of pathways is a major focus of present-day biology. Typical pathways involve large numbers of entities of various types whose associations are represented as reactions involving arbitrary numbers of reactants, outputs and modifiers. Until recently, few information extraction approaches were capable of resolving the level of detail in text required to support the annotation of such pathway representations. We argue that event representations of the type popularized by the BioNLP Shared Task are potentially applicable for pathway annotation support. As a step toward realizing this possibility, we study the mapping from a formal pathway representation to the event representation in order to identify remaining challenges in event extraction for pathway annotation support. Following initial analysis, we present a detailed study of protein association and dissociation reactions, proposing a new event class and representation for the latter and, as a step toward its automatic extraction, introduce a manually annotated resource incorporating the type among a total of nearly 1300 annotated event instances. As a further practical contribution, we introduce the first pathway-to-event conversion software for SBML/CellDesigner pathways and discuss the opportunities arising from the ability to convert the substantial existing pathway resources to events.

1 Introduction

For most of the previous decade of biomedical information extraction (IE), efforts have focused on foundational tasks such as named entity detection and their database normalization (Krallinger et al., 2008) and simple IE targets, most commonly binary entity relations representing associations such as protein-protein interactions (Pyysalo et al., 2008; Tikk et al., 2010). In recent years, an increasing number of resources and methods pursuing more detailed representations of extracted information are becoming available (Pyysalo et al., 2007; Kim et al., 2008; Thompson et al., 2009; Björne et al., 2010). The main thrust of this move toward structured, fine-grained information extraction falls under the heading of *event extraction* (Ananiadou et al., 2010), an approach popularized and represented in particular by the BioNLP Shared Task (BioNLP ST) (Kim et al., 2009a; Kim et al., 2011).

While a detailed representation of extracted information on biomolecular events has several potential applications ranging from semantic search to database curation support (Ananiadou et al., 2010), the number of practical applications making use of this technology has arguably so far been rather limited. In this study, we pursue in particular the opportunities that event extraction holds for pathway annotation support,[1] arguing that the match between

[1]Throughout this paper, we call the projected task *pathway annotation support*. There is no established task with this label, and we do not envision this to be a specific single task. Rather, we intend the term to refer to a set of tasks where information extraction/text mining methods are applied in some role to contribute directly to pathway curation, including, for example, the identification of specific texts in the literature relevant to annotated reactions, the automatic suggestion of further entities or reactions to add to a pathway, or even the fully automatic generation of entire pathways from scratch.

Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011, pages 105–113,
Portland, Oregon, USA, June 23-24, 2011.

representations that biologists employ to capture reactions between biomolecules in pathways and the event representation of the BioNLP ST task makes pathway-oriented applications a potential "killer application" for event extraction technology.

The fit between these representations is not accidental – the design of the BioNLP ST event representation has been informed by that of popular pathway models – nor is it novel to suggest to support pathway extraction through information methods in general (see e.g. (Rzhetsky et al., 2004)) or through event extraction specifically (Oda et al., 2008). However, our study differs from previous efforts in two key aspects. First, instead of being driven by information extraction and defining a representation fitting its results, we specifically adopt the perspective and model of a widely applied standard database representation and proceed from the pathway to events in text. Second, while previous work on event extraction for pathway annotation has been exploratory in nature or has otherwise had limited practical impact, we introduce and release a first software implementation of a conversion from a standard pathway format to the event format, thus making a large amount of pathway data available for use in event extraction and taking a concrete step toward reliable, routine mappings between the two representations.

2 Representations and Resources

Before proceeding to consider the mapping between the two, we first briefly introduce the pathway and event representations in focus in this study and the applied pathway resources.

2.1 Pathways

The biomolecular curation community has created and made available an enormous amount of pathway resources: for example, as of April 2011, the Pathguide pathway resource list[2] includes references to 325 pathway-related resources – many of which are themselves pathway databases containing hundreds of individual models. These resources involve a formidable variety of different, largely independently developed formats and representations of which only few pairs have tools supporting mutual conversion. To address the challenges of interoperability that this diversity implies, a number of standardization efforts for pathway representations have been introduced.

In this work, we consider two widely adopted pathway representation formats: Systems Biology Markup Language (SBML)[3] (Hucka et al., 2003) and Biological Pathway Exchange (BioPAX)[4] (Demir et al., 2010). SBML is an XML-based machine-readable data exchange format that supports a formal mathematical representation of chemical reactions (including e.g. kinetic parameters), allowing biochemical simulation. BioPAX is an RDF/OWL-based standard language to represent bio-molecular and cellular networks designed to enable data integration, exchange, visualization and analysis. Despite significantly different choices in storage format, the represented information content of the two is broadly compatible. In the following, we refer to established correspondences and mappings when relating the two (see e.g. (Mi and Thomas, 2009)).

As an interchange format aimed to support a large variety of specific representations, the SBML standard itself does not define a fixed set of types of physical entities or biochemical reactions. However, the standard defines an extension mechanism allowing additional information, including such types, to be defined. As specific, fixed types with established semantics are a requirement for practical conversion between the different representations, we thus rely in this work not only on SBML core, but also a minimal set of the extensions introduced by the popular CellDesigner pathway modeling tool (Funahashi et al., 2008). In the following, we assume throughout the availability of CellDesigner extensions when discussing SBML features.

For pathway data, in this study we use the full set of pathways contained in the Panther and Payao pathway repositories in SBML form. Panther (Protein ANalysis THrough Evolutionary Relationships) is a gene function-based classification system that hosts a large collection of pathways. The Panther repository consists of 165 pathways, including 153 signaling and 12 metabolic pathways. All pathways

[2] http://www.pathguide.org/

[3] http://sbml.org

[4] http://www.biopax.org

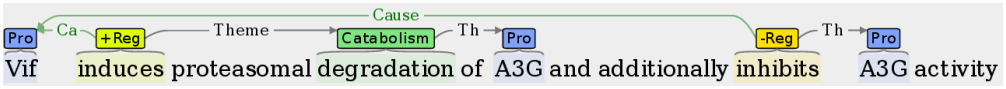

Figure 1: Illustration of the event representation.

were drawn on CellDesigner by manual curation and thus include CellDesigner SBML extensions (Mi and Thomas, 2009). Payao is a community-based SBML model tagging platform (Matsuoka et al., 2010) that allows a community to share models, tag and add comments, and search relevant literature (Kemper et al., 2010). Currently, 28 models are registered in Payao. As in Panther, all Payao pathways include CellDesigner extensions.

2.2 Event Representation

The application of *event representations* in biomedical IE is a relatively recent development, following the introduction of corpus resources annotating structured, n-ary associations of entities with detailed types (Pyysalo et al., 2007; Kim et al., 2008; Thompson et al., 2009)) and popularized in particular by the BioNLP Shared Task (BioNLP ST) events (Kim et al., 2009b; Kim et al., 2011). In this paper, we use *event* in the BioNLP ST sense, to refer specifically to the representation where each event is assigned a type from a fixed ontology, bound to a specific expression in text stating its occurrence (the *trigger* or *text binding*), and associated with an arbitrary number of participants (similarly text-bound entities or other events), for which the roles in which they are involved in the event are defined from a fixed small inventory of event argument types (e.g. *Theme*, *Cause*, *Site*). These concepts are illustrated in Figure 1.

3 Analysis of Pathway-Event Mapping

We next present an analysis of the relationship between the two representations, considering features required from IE systems for efficient support of pathway annotation support.

We assume throughout that the target on the pathway side is restricted to the broad, central biological content of pathways, excluding information only related to e.g. simulation support or pathway visualization/layout.

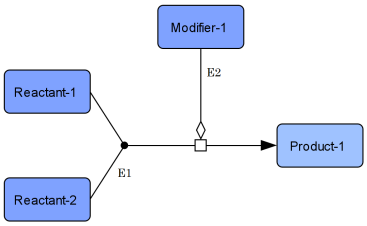

Figure 2: Illustration of a generalized pathway reaction.

3.1 Top-level concepts

Both SBML and BioPAX involve two (largely comparable) top-level concepts that form the core of the representation: entity (species/physical entity) and reaction (interaction). In the following we focus primarily on entities and reactions, deferring consideration of detailed concepts such as modification state and compartment localization to Section 3.3.

The concept of a reaction in the considered pathway representations centrally involves three sets of entities: reactants, products, and modifiers. As the names suggest, the reaction produces the set of product entities from the reactant entities and is affected by the modifiers. Figure 2 shows an illustration of a generalized reaction. Pathway reactions find a reasonably good analogy in events in the event representation. While the event representation does not differentiate "reactants" from "products" in these terms, the roles assigned to event participants allow comparable interpretation. There is no single concept in the event representation directly comparable to reaction modifiers. However, the semantics of specific modification types (see Section 3.3) correspond broadly to those of regulation in the event representation, suggesting that modification be represented using a separate event of the appropriate type with the modifying entities participating in the *Cause* role (Kim et al., 2008). Figure 3 illustrates the event structure proposed to correspond to the reaction of Figure 2, with the added assumptions that the reaction and modification types (unspecified in Figure 2) are Association (BioPAX:ComplexAssembly) and Modulation (BioPAX:Control).

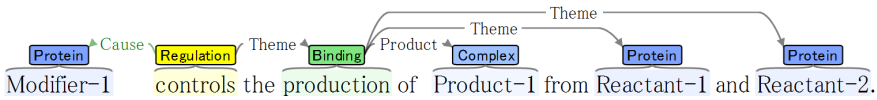

Figure 3: Illustration of a generalized event structure with four entities and two events (REGULATION and BINDING). Note that the text is only present as filler to satisfy the requirement that events are bound to specific expressions in text. The *Product* role is not a standard role in event representation but newly proposed in this study.

Pathway		Event		
CellDesigner	BioPAX	ST'09	ST'11	GENIA
Protein	Protein	Protein	Protein	Protein
RNA	RNA	Protein	Protein	RNA
AntiSenseRNA	RNA	Protein	Protein	RNA
Gene	DNA	Protein	Protein	DNA
Simple molecule	Small molecule	-	Chemical	Inorganic compound
Ion	Small molecule	-	Chemical	Inorganic compound
Drug	PhysicalEntity	-	Chemical	Inorganic compound
Hetero/homodimer	Complex	-	-	Protein complex

Table 1: Entity type comparison between pathways and events.

The mapping of top-level concepts that we consider thus unifies physical entities in pathways with the entities of the BioNLP ST representation, and pathway *reaction* with *event*.[5]

To be able to efficiently support (some aspect of) pathway annotation through IE, the applied extraction model should be able, for both entities and reactions, to 1) recognize mentions of all relevant types of entity/reaction and 2) differentiate between entity/reaction types at the same or finer granularity as the pathway representation. For example, an IE system that does not detect mentions of protein complexes cannot efficiently support aspects of pathway annotation that involve this type; a system that detects proteins and complexes with no distinction between the two will be similarly limited. In the following, we consider entity and reaction types separately to determine to what extent these requirements are filled by presently available resources for event extraction, in particular the GENIA corpus (Kim et al., 2008) and the BioNLP ST 2009 (Kim et al., 2009b) and 2011 corpora.

[5]Pathways and IE/text mining use many of the same terms with (sometimes subtly) different meanings. We use largely IE terminology, using e.g. *entity* instead of *species* (SBML) and *entity type* instead of *physical entity class* (BioPAX) / *species type* (SBML) For the pathway associations, we have adopted *reaction* (SBML term) in favor of *interaction* (BioPAX). With *event*, we refer to the BioNLP ST sense of the word; we make no use of the SBML "event" concept.

3.2 Entities

Table 1 shows a comparison of the primary entity types between SBML/CellDesigner, BioPAX, and the event representations. There is significant difference in the resolution of gene and gene product types between the pathway representations and that applied in ST'09 and ST'11: while both pathway representations and the GENIA corpus differentiate the DNA, RNA and protein forms, the STs fold the three types into a single one, PROTEIN.[6] The CHEMICAL type defined in ST'11 (ID task) overlaps largely with BioPAX SMALL MOLECULE, a type that SBML/CellDesigner further splits into two specific types, and further partly covers the definition of the SBML/CellDesigner type Drug. The same holds (with somewhat less specificity) for GENIA INORGANIC COMPOUND. Finally, although annotated in GENIA, the category of protein complexes has no correspondence among the entities considered in the BioNLP ST representation.

Thus, information extraction systems applying the core BioNLP ST entity types will entirely lack coverage for protein complexes and will not be able

[6]While the term PROTEIN appears to suggest that the class consists only of protein forms, these entities are in fact annotated in the BioNLP ST data according to the GENIA gene/gene product guidelines (Ohta et al., 2009) and thus include also DNA and RNA forms. The type could arguably more accurately be named GENE OR GENE PRODUCT.

Pathway		Event		
CellDesigner	BioPAX	ST'09	ST'11	GENIA
State transition	BiochemicalReaction	(see Table 3)		
Truncation	BiochemicalReaction	Catabolism	Catabolism	Catabolism
Transcription	BiochemicalReaction	Transcription	Transcription	Transcription
Translation	BiochemicalReaction	-	-	Translation
Association	ComplexAssembly	Binding	Binding	Binding
Dissociation	ComplexAssembly	-	-	-
Transport	Transport w/reaction	Localization	Localization	Localization
Degradation	Degradation	Catabolism	Catabolism	Catabolism
Catalysis	Catalysis	Positive regulation	Positive regulation	Positive regulation
Physical stimulation	Control	Positive regulation	Positive regulation	Positive regulation
Modulation	Control	Regulation	Regulation	Regulation
Trigger	Control	Positive regulation	Positive regulation	Positive regulation
Inhibition	Control	Negative regulation	Negative regulation	Negative regulation

Table 2: Reaction type comparison between pathways and events.

to fully resolve the detailed type of gene and gene product types applied in the pathway representations. While these distinctions exist in the full GENIA corpus, it has not been frequently applied in event extraction in its complete form and is unlikely to be adopted over the widely applied ST resources. Finally, none of the event representations differentiate the pathway small molecule/drug types. We discuss the implications of these ambiguities in detail below. By contrast, we note that both SBML/CellDesigner and BioPAX entity types cover the scope of the major BioNLP ST types and have comparable or finer granularity in each case.

3.3 Reactions

Table 2 shows a comparison between the reaction types of the two considered pathway representations and those of the BioNLP ST event representation. The full semantics of the generic reaction type State transition (BioPAX: BiochemicalReaction) cannot be resolved from the type alone; we defer discussion of this type.

Contrary to the event types, we find that for reaction types even the least comprehensive BioNLP ST'09 event representation has high coverage of the pathway reaction types as well as a largely comparable level of granularity in its types. While neither of the BioNLP ST models defines a TRANSLATION type, the adoption of the GENIA representation – matching that for TRANSCRIPTION – for this simple and relatively rare event type would likely be relatively straightforward. A more substantial omission in all of the event representations is the lack of a Dissociation event type. As dissociation is the "reverse" reaction of (protein) BINDING and central to many pathways, its omission from the event model is both surprising as well as potentially limiting for applications of event extraction to pathway annotation support.

The detailed resolution of pathway reactions provided by the event types has implications on the impact of the ambiguity noted between the single type covering genes and gene products in the event representation as opposed to the distinct DNA/RNA/protein types applied in the pathways. Arguably, for many practical cases the specific type of an entity of the broad gene/gene product type is unambiguously resolved by the events it participates in: for example, any gene/gene product that is modified through phosphorylation (or similar reaction) is necessarily a protein.[7] Similarly, only proteins will be involved in e.g. localization between nucleus and cytoplasm. On a more detailed level, BINDING events resolves their arguments in part through their *Site* argument: binding to a promoter implies DNA, while binding to a C-terminus implies protein. Thus, we can (with some reservation) forward the argument that it is not necessary to disambiguate all gene/gene product mentions on the entity level for pathway annotation support, and that successful event extraction will provide disambiguation in cases where the distinction matters.

[7] DNA methylation notwithstanding; the BioNLP ST'11 EPI task demonstrated that protein and DNA methylation can be disambiguated on the event type level without entity type distinctions.

Pathway			Event		
SBML/CellDesigner			ST'09	ST'11	GENIA
in:Compartment$_1$	→	in:Compartment$_2$	Localization	Localization	Localization
residue:state:∅	→	residue:state:Phosphorylated	Phosphorylation	Phosphorylation	Phosphorylation
residue:state:Phosphorylated	→	residue:state:∅	-	Dephosphorylation	Dephosphorylation
residue:state:∅	→	residue:state:Methylated	-	Methylation	Methylation
residue:state:Methylated	→	residue:state:∅	-	Demethylation	Demethylation
residue:state:∅	→	residue:state:Ubiquitinated	-	Ubiquitination	Ubiquitination
residue:state:Ubiquitinated	→	residue:state:∅	-	Deubiquitination	Deubiquitination
species:state:inactive	→	species:state:active	Positive regulation	Positive regulation	Positive regulation
species:state:active	→	species:state:inactive	Negative regulation	Negative regulation	Negative regulation

Table 3: Interpretation and comparison of state transitions.

Finally, the pathway representations define generic reaction types (State transition/BiochemicalReaction) that do not alone have specific interpretations. To resolve the event involved in these reactions it is necessary to compare the state of the reactants against that of the matching products. Table 3 shows how specific state transitions map to event types (this detailed comparison was performed only for SBML/CellDesigner pathways). We find here a good correspondence for transitions affecting a single aspect of entity state. While generic pathway transitions can change any number of such aspects, we suggest that decomposition into events where one event corresponds to one point change in state is a reasonable approximation of the biological interpretation: for example, a reaction changing one residue state into Methylated and another into Phosphorylated would map into two events, METHYLATION and PHOSPHORYLATION.

In summary of the preceding comparison of the core pathway and event representations, we found that in addition to additional ambiguity in e.g. gene and gene product types, the popular BioNLP ST representations lack a protein complex type and further that none of the considered event models define a (protein) dissociation event. To address these latter omissions, we present in the following section a case study of dissociation reactions as a step toward their automatic extraction. We further noted that pathway types cover the event types well and have similar or higher granularity in nearly all instances. This suggests to us that mapping from the pathway representation to events is more straightforward than vice versa. To follow up on these opportunities, we introduce such a mapping in Section 5, in following the correspondences outlined above.

4 Protein Association and Dissociation

In the analysis presented above, we noted a major reaction type defined in both considered pathway representations that had no equivalent in the event representation: dissociation. In this section, we present a study of this reaction type and its expression as statements in text through the creation of event-style annotation for dissociation statements.

4.1 Target data

Among the large set of pathways available, we chose to focus on the Payao mTOR pathway (Caron et al., 2010) because it is a large, recently introduced pathway with high-quality annotations that involves numerous dissociation reactions. The Payao pathways are further annotated with detailed literature references, providing a PubMed citation for nearly each entity and reaction in the pathway. To acquire texts for event annotation, we followed the references in the pathway annotation and retrieved the full set of PubMed abstracts associated with the pathway, over 400 in total. We then annotated 60 of these abstracts that were either marked as relevant to dissociation events in the pathway or were found to include dissociation statements in manual analysis. These abstracts were not included in any previously annotated domain corpus. Further, as we aimed specifically to be able to identify event structures for which no previous annotations exist, we could not rely on (initial) automatic annotation.

4.2 Annotation guidelines

We performed exhaustive manual entity and event annotation in the event representation for the selected 60 abstracts. For entity annotation, we ini-

tially considered adopting the gene/gene product annotation guidelines (Ohta et al., 2009) applied in the BioNLP ST 2009 as well as in the majority of the 2011 tasks. However, the requirement of these guidelines to mark only specific gene/protein names would exclude a substantial number of the entities marked in the pathway, as many refer to gene/protein families or groups instead of specific individual genes or proteins. We thus chose to adopt the pathway annotation itself for defining the scope of our entity annotation: we generated a listing of all the names appearing in the target pathway and annotated their mentions, extrapolating from this rich set of examples to guide us in decisions on how to annotate references to entities not appearing in the pathway. For event annotation, we adapted the GENIA event corpus annotation guidelines (Kim et al., 2008), further developing a specific representation and guidelines for annotating dissociation events based on an early iteration of exploratory annotation.

Annotation was performed by a single biology PhD with extensive experience in event annotation (TO). While we could thus not directly assess inter-annotator consistency, we note that our recent comparable efforts have been evaluated by comparing independently created annotations at approximately 90% F-score for entity annotations and approximately 80% F-score for event annotations (BioNLP Shared Task primary evaluation criteria) (Pyysalo et al., 2011; Ohta et al., 2011).

4.3 Representing Association and Dissociation

Based on our analysis of 107 protein dissociation statements annotated in the corpus and a corresponding study of the "reverse", statements of protein association in the corpus, we propose the following extensions for the BioNLP ST event representation. First, the introduction of the event type DISSOCIATION, taking as its primary argument a single *Theme* identifying a participating entity of the type COMPLEX. Second, we propose the new role type *Product*, in the annotation of DISSOCIATION events an optional (secondary) argument identifying the PROTEIN entities that are released in the dissociation event. This argument should be annotated (or extracted) only when explicitly stated in text. Third, for symmetry in the representation, more detail in extracted information, and to have a representation

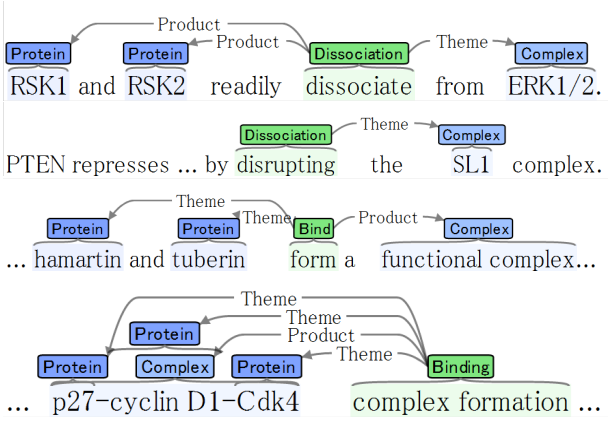

Figure 4: Examples annotated with the proposed event representation for DISSOCIATION and BINDING events with the proposed *Product* role marking formed complex.

Item	Count
Abstract	60
Word	11960
Protein	1483
Complex	201
Event	1284

Table 4: Annotation statistics.

more compatible with the pathway representation for protein associations, we propose to extend the representation for BINDING, adding *Product* as an optional argument identifying a COMPLEX participant in BINDING events marking statements of complex formation stating the complex. The extended event representations are illustrated in Figure 4.

4.4 Annotation statistics

Table 4 presents the statistics of the created annotation. While covering a relatively modest number of abstracts, the annotation density is very high, relating perhaps in part to the fact that many of the referenced documents are reviews condensing a wealth of information into the abstract.

5 Pathway-to-event conversion

As an additional practical contribution and outcome of our analysis of the mapping from the pathway representation to the event representation, we created software implementing this mapping from SBML with CellDesigner extensions to the event representation. This conversion otherwise follows

the conventions of the event model, but lacks specific text bindings for the mentioned entities and event expressions (triggers). To maximize the applicability of the conversion, we chose to forgo e.g. the CellDesigner plugin architecture and to instead create an entirely standalone software based on python and libxml2. We tested this conversion on the 165 Panther and 28 Payao pathways to assure its robustness.

Conversion from pathways into the event representation opens up a number of opportunities, such as the ability to directly query large-scale event repositories (e.g. (Björne et al., 2010)) for specific pathway reactions. For pathways where reactions are marked with literature references, conversion further allows event annotations relevant to specific documents to be created automatically, sparing manual annotation costs. While such event annotations will not be bound to specific text expressions, they could be used through the application of techniques such as distant supervision (Mintz et al., 2009). As a first attempt, the conversion introduced in this work is limited in a number of ways, but we hope it can serve as a starting point for both wider adoption of pathway resources for event extraction and further research into accurate conversions between the two. The conversion software, `SBML-to-event`, is freely available for research purposes.

6 Discussion and Conclusions

Over the last decade, the bio-community has invested enormous efforts in the construction of detailed models of the function of a large variety of biological systems in the form of pathways. These efforts toward building systemic understanding of the functioning of organisms remain a central focus of present-day biology, and their support through information extraction and text mining perhaps the greatest potential contribution that the biomedical natural language processing community could make toward the broader bio-community.

We have argued that while recent developments in BioNLP are highly promising for approaching practical support of pathway annotation through information extraction, the BioNLP community has not yet made the most of the substantial resources in the form of existing pathways and that pursuing mapping from pathways to the event representation might be both more realistic and more fruitful than the other way around. As a first step in what we hope will lead to broadened understanding of the different perspectives, communication between the communities, and better uses resources, we have introduced a fully automatic mapping from SBML/CellDesigner pathways into the BioNLP ST-style event representation. As a first effort this mapping has many limitations and imperfections that we hope the BioNLP community will take as a challenge to do better.

Noting in analysis that dissociation reactions are not covered in previously proposed event representations, we also presented a detailed case study focusing on statements describing protein association and dissociation reactions in PubMed abstracts relevant to the mTOR pathway. Based on exploratory annotation, we proposed a novel event class DISSOCIATION, thus taking a step toward covering this arguably most significant omission in the event representation.

The pathway-bound event annotations created in this study, exhaustive annotation of all relevant entities and events in 60 abstracts, consist in total of annotation identifying nearly 1500 protein and 200 complex mentions and over 1200 events involving these entities in text. These annotations are freely available for use in research at `http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA`.

Acknowledgments

We would like to thank Hiroaki Kitano, Yukiko Matsuoka and Samik Ghosh of the Systems Biology Institute for their generosity in providing their time and expertise in helping us understand the CellDesigner and SBML pathway representations. This work was partially supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan).

References

Sophia Ananiadou, Sampo Pyysalo, Jun'ichi Tsujii, and Douglas B. Kell. 2010. Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*, 28(7):381–390.

Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. 2010. Complex event extraction at PubMed scale. *Bioinformatics*, 26(12):i382–390.

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. *Molecular Systems Biology*, 6(1).

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. *Nature biotechnology*, 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. *Proceedings of the IEEE*, 96(8):1254–1265.

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.

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.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(10).

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009a. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of BioNLP 2009 Shared Task*.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009b. Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop*, pages 1–9.

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 2011*.

M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirschman, and A. Valencia. 2008. Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome biology*, 9(Suppl 2):S1.

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

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

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL'09*, pages 1003–1011.

Kanae Oda, Jin-Dong Kim, Tomoko Ohta, Daisuke Okanohara, Takuya Matsuzaki, Yuka Tateisi, and Jun'ichi Tsujii. 2008. New challenges for text mining: Mapping between text and manually curated pathways. *BMC Bioinformatics*, 9(Suppl 3):S5.

Tomoko Ohta, Jin-Dong Kim, Sampo Pyysalo, Yue Wang, and Jun'ichi Tsujii. 2009. Incorporating GENETAG-style annotation to GENIA corpus. In *Proceedings of BioNLP'09*, pages 106–107.

Tomoko Ohta, Sampo Pyysalo, and Jun'ichi Tsujii. 2011. Overview of the Epigenetics and Post-translational Modifications (EPI) task of BioNLP Shared Task 2011. In *Proceedings of BioNLP 2011*.

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(50).

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

Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Dan Sullivan, Chunhong Mao, Chunxia Wang, Bruno Sobral, Jun'ichi Tsujii, and Sophia Ananiadou. 2011. Overview of the Infectious Diseases (ID) task of BioNLP Shared Task 2011. In *Proceedings of BioNLP 2011*.

Andrey Rzhetsky, Ivan Iossifov, Tomohiro Koike, Michael Krauthammer, Pauline Kra, Mitzi Morris, Hong Yu, Pablo Ariel Duboué, Wubin Weng, W. John Wilbur, Vasileios Hatzivassiloglou, and Carol Friedman. 2004. GeneWays: A system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, 37(1):43–53.

Paul Thompson, Syed Iqbal, John McNaught, and Sophia Ananiadou. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10(1):349.

Domonkos Tikk, Philippe Thomas, Peter Palaga, Jörg Hakenberg, and Ulf Leser. 2010. A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Comput Biol*, 6(7):e1000837, 07.