Pathway Curation (PC) task at BioNLP Shared Task 2013 Task Proposal

Sampo Pyysalo, Tomoko Ohta and Sophia Ananiadou

National Centre for Computer Science (NaCTeM) and
School of Computer Science, University of Manchester, UK

Task Summary

Task name Pathway Curation (PC)

Goal in IE To evaluate and advance event extraction methods for the capacity to effectively support pathway model curation tasks.

Motivation in biology To support the curation, evaluation and maintenance of biomolecular pathway models.

Description of the entities and events Entities: simple chemicals, genes and gene products, complexes, and cellular components. Events: modifications of covalent (e.g. phosphorylation) and noncovalent (e.g. binding) bonds, synthesis, degradation and transport reactions, and regulatory relationships. Both entities and events are defined with respect to the Systems Biology Ontology and the Gene Ontology, as detailed below.

Concrete examples in BioNLP format Available in attached data.

Evaluation and criteria Gold entity annotations are provided as input to participants, events and their modifications (Negation and Speculation) need to be predicted. Evaluation is mention-based and follows the BioNLP ST'09 [10] and ST'11 GE, EPI and ID criteria [9, 16].

Task corpus PubMed abstracts selected by relevance to specific reactions in pathways of interest to task organizers, including both signaling and metabolic pathways. The ongoing annotation effort aims for 800-1000 annotated documents in total, expected to involve at least 10,000 entity and 10,000 event annotations.

Existing corpus mTOR pathway corpus [15], containing approximately 1,700 entity and 1,300 event annotations.

Feasibility of task The task combines aspects of the BioNLP ST'09 task and the BioNLP ST'09 GE and EPI tasks, with extensions specific to the target. mTOR pathway corpus extraction has been evaluated using EventMine (manuscript in review), giving an F-score of over 50% using standard BioNLP ST evaluation criteria.

Annotation process Event annotation is created by six annotators lead by one senior annotator, all PhD-level biologists, three with a background in signaling and four in metabolic pathways. Inter-annotator agreement is measured throughout the annotation process by random double annotation of a sample of documents, conflicts resolved by discussion with lead annotator.

Tentative schedule for annotated corpus delivery The annotation effort is aimed to complete in the latter half of November 2012 with a random sample delivered as training data to task participants on December 1st, 2012.

Contents

1 Introduction

This document presents the Pathway Curation (PC) task proposed for the BioNLP Shared Task 2013.

The PC task aims to evaluate the applicability of event extraction systems to support the curation, evaluation and maintenance of biomolecular pathway models and to encourage the further development of methods for these tasks.

Despite more than a decade of work in biomedical text mining on tasks under headings such as "automatic pathway extraction", natural language processing and information extraction methods have not been widely embraced by biomedical pathway curation communities. Until recently, biomedical domain IE efforts concentrated on simple representations (e.g. physical entity pairs) that were not sufficiently expressive to address pathway curation, and most work also involved different semantics from those applied in curation efforts. We believe that the structured event representation applied in BioNLP Shared Task main tasks offers many opportunities to make a significant contribution to practical pathway curation efforts, and propose the PC task as a step toward realizing these opportunities.

2 Representation and Task Setting

The BioNLP Shared Task (ST) 2013 Pathway Curation (PC) task is an event extraction task following the representation and task setting of the ST'09 and ST'11 main tasks.

The representation involves two primary categories of annotation: (physical) entity annotation, and event annotation. Mentions of physical entities in text are annotated as contiguous spans that are assigned types (e.g. GENE OR GENE PRODUCT, CELLULAR COMPONENT). References to reactions, processes and other similar associations are annotated using the *event representation*, where each event annotation (*event* for short) is associated with a specific contiguous span in text (the *event trigger*), assigned a type (e.g. PHOSPHORYLATION), and associated with its *participants* (entities or other events), each of which is marked as participating in the event in a specific *role* (e.g. *Theme*, *Cause*). Figure 1 shows illustrations of some annotations.

In the shared task, participants in the PC task will be provided with *gold standard* annotations for entity mentions. The task thus focuses participants' efforts on the primary event extraction task. This task setting follows that of main tasks in ST'09 and ST'11 and will thus be familiar to participants who have taken part in these events and compatible with existing systems introduced to address these tasks. Additionally, this task setting allows the application of the evaluation crite-

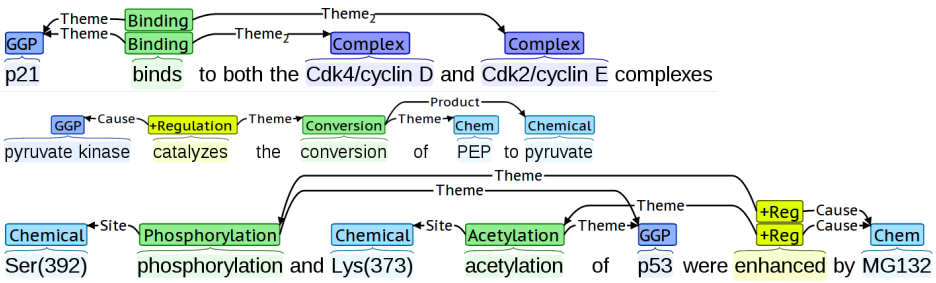

Figure 1: Illustrations of event annotations. +REG and -REGULATION abbreviate for POSITIVE REGULATION and NEGATIVE REGULATION, respectively.

ria and tools introduced in the previous shared tasks, thus assuring that results are broadly comparable across tasks and removing the need to define and implement novel evaluation criteria.

3 Document Selection

The corpus texts are selected on the basis of relevance to a selection of pathway models from Panther Pathway DB [13] and BioModels [11], covering both signaling and metabolic pathways.

Specifically, a set of candidate reactions were first selected at random from a set of pathways of interest using the PathText semantic search system integrator [8]. The PathText system then returns a set of candidate documents for each reaction ranked by relevance. For each reaction returning at least 10 candidate documents, the actual relevance of these documents to the reaction was then manually evaluated in a separate annotation task preceding the event annotation, in collaboration between NaCTeM/University of Manchester and the Korea Institute of Science and Technology Information (KISTI).

The PC event annotation task documents are then selected as a random subset from among those judged relevant to reactions from pathways of interest in this process. The final size of this random sample of relevant documents will be decided according to the resources available for the manual annotation effort, with 800-1000 documents as initial target.

4 Entity Annotation

This section introduces the entity annotation applied in the PC task.

4.1 Annotation overview

The entity annotation marks mentions of selected types of entities in text. An entity mention annotation is defined by start and end offsets (contiguous span) in text and a type selected from a fixed set. The following types are annotated for the PC task:

Type	Scope	Reference
SIMPLE CHEMICAL	simple, non-repetitive chemical entities	ChEBI
GENE OR GENE PRODUCT	genes, RNA and proteins	gene/protein DBs
COMPLEX	entities of non-covalently linked components	complex DBs
CELLULAR COMPONENT	parts of cell and extracellular environment	GO-CC

GENE OR GENE PRODUCT and SIMPLE CHEMICAL mentions have been annotated also in numerous other tasks in previous shared tasks in the BioNLP ST series, and are annotated using similar guidelines. SIMPLE CHEMICAL is annotated with reference to the Chemical Entities of Biological Interest (ChEBI) resource [5], and GENE OR GENE PRODUCT is annotated with reference to gene and protein databases such as UniProt [2], Entrez Gene [12] and PFam [6]. COMPLEX, marking mentions of named macromolecular complexes, will be similarly annotated with reference to database resources covering complexes. CELLULAR COMPONENT is annotated following the scope of the Gene Ontology `cellular component` subontology [1]. For discussion of the relation between these types and the representations applied in pathway models, we refer to [15].

In terms of mention types in text, the annotation for SIMPLE CHEMICAL, GENE OR GENE PRODUCT and COMPLEX covers entity name mentions only, while the annotation for CELLULAR COMPONENT covers entity name mentions, nominal mentions, and adjectival mentions (e.g. "*mitochondrial*"). Further details are given in the task annotation guidelines.

4.2 Annotation process

The base entity annotation for the PC corpus will be created automatically using state-of-the-art entity mention taggers for each of the targeted entity types, integrated in the Argo workflow system [17].

To focus resources on the creation of event annotations, there is no separate manual annotation stage for correcting the base entity annotation. Annotators should correct errors in the automatically created entity annotation while creating event annotation, but the quality of the entity annotation will not be separately assessed.

The following tools and resources will be used to create the base entity annotation:

Type	Tagger
SIMPLE CHEMICAL	OSCAR4
GENE OR GENE PRODUCT	NERsuite
COMPLEX	Dictionary and heuristics
CELLULAR COMPONENT	NERsuite

For SIMPLE CHEMICAL tagging, the OSCAR4 system [7] trained on the chemical named entity recognition corpus of [3] will be applied. For GENE OR GENE PRODUCT mention detection, the NERsuite[1] system trained on the BioCreative 2 Gene Mention task [19] corpus will be used. NERsuite will also be applied for CELLULAR COMPONENT mention detection, for this task trained on the Anatomical Entity Mention (AnEM) corpus [14]. Finally, COMPLEX annotations will be created using a combination of a dictionary and heuristics making use of the GENE OR GENE PRODUCT annotation (for mentions such as "*cyclin E/CDK2 complex*").

Based on evaluations of each of these tools, we estimate that the automatic tagging performance will fall in the range 80-90% in both precision and recall and expected to reach 90% through revision during event annotation. This level of performance is estimated to be sufficient for the task .[2]

[1] http://nersuite.nlplab.org/

[2] The entity annotation serves as the basis of the event annotation, but it is not itself a target of extraction for shared task participants as gold entity annotations will be provided as part of input. For this reason, while quality is important, it is not a primary concern for the use of the corpus that the entity annotation quality be as high as possible and that the marking is without any bias, and automatic methods can thus be applied to support the entity annotation task.

5 Event annotation

This section introduces the event annotation applied in the PC task.

5.1 Annotation overview

The event annotation marks references to reactions, processes and comparable associations that are in scope of the annotation using the event representation (Section 2). The following events are annotated for the PC task:

Event type	Arguments	Reference
CONVERSION	*Theme*:Molecule, Product:Molecule	SBO
PHOSPHORYLATION	*Theme*:Molecule *Site*:SIMPLE CHEMICAL	SBO
DEPHOSPHORYLATION	*Theme*:Molecule *Site*:SIMPLE CHEMICAL	SBO
ACETYLATION	*Theme*:Molecule *Site*:SIMPLE CHEMICAL	SBO
DEACETYLATION	*Theme*:Molecule *Site*:SIMPLE CHEMICAL	SBO
(etc.)	*Theme*:Molecule *Site*:SIMPLE CHEMICAL	SBO
LOCALIZATION	*Theme*:Molecule, *At,From,To*:CELLULAR COMPONENT	GO
TRANSPORT	*Theme*:Molecule, *From,To*:CELLULAR COMPONENT	SBO
GENE EXPRESSION	*Theme*:GENE OR GENE PRODUCT	GO
TRANSCRIPTION	*Theme*:GENE OR GENE PRODUCT	SBO
TRANSLATION	*Theme*:GENE OR GENE PRODUCT	SBO
DEGRADATION	*Theme*:Molecule	SBO
BINDING	*Theme*:Molecule, Product:COMPLEX	SBO
DISSOCIATION	*Theme*:COMPLEX, Product:Molecule	SBO
REGULATION	*Theme*:ANY, *Cause*:ANY	GO
POSITIVE REGULATION	*Theme*:ANY, *Cause*:ANY	GO
ACTIVATION	*Theme*:Molecule, *Cause*:ANY	GO
NEGATIVE REGULATION	*Theme*:ANY, *Cause*:ANY	GO
INACTIVATION	*Theme*:Molecule, *Cause*:ANY	GO
PATHWAY	*Participant*:Molecule	SBO

Here, "Molecule" refers to an entity annotation of any of the types SIMPLE CHEMICAL, GENE OR GENE PRODUCT, or COMPLEX, and "ANY" refers to an annotation of any type, either entity or event. The indentation corresponds to ontological relationships between the event types.[3]

For the definition and scope of the event annotation, we rely primarily on the Systems Biology Ontology (SBO) [4], drawing some general types not in scope of this ontology from the Gene Ontology (GO) [1]. The definitions of these types are summarized in Section 5.2 and the arguments in Section 5.3.

[3]For example, PHOSPHORYLATION `is-a` CONVERSION and TRANSCRIPTION `part-of` GENE EXPRESSION.

5.2 Annotated type definitions

The following definitions of the annotated event types are primarily drawn the corresponding definitions in SBO and GO, with minor modifications (mostly for length).

Type	Definition
CONVERSION	Biochemical reaction that results in the modification of some covalent bonds.
PHOSPHORYLATION	Addition of a phosphate group (-H2PO4) to an entity.
DEPHOSPHORYLATION	Removal of a phosphate group (-H2PO4) from an entity.
(etc.)	Addition/removal of a chemical group to/from an entity.
LOCALIZATION	Process in which an entity is transported to or maintained in a specific location.
TRANSPORT	Movement of a physical entity without modification of the structure of the entity.
GENE EXPRESSION	Process in which a gene sequence is converted into a mature gene product or products
TRANSCRIPTION	Process through which a DNA sequence is copied to produce a complementary RNA.
TRANSLATION	Process in which a polypeptide chain is produced from a messenger RNA.
DEGRADATION	Complete disappearance of a physical entity.
BINDING	Interaction between entities that results in the formation of a non-covalent complex.
DISSOCIATION	Transformation of complex that results in the formation of several independent entities.
REGULATION	Any process that modulates any attribute of any process, quality or function.
POSITIVE REGULATION	Process that activates or increases the frequency, rate or extent of a process.
ACTIVATION	Process that results in the transition of an entity from an inactive to an active state.
NEGATIVE REGULATION	Process that stops, prevents, or reduces the frequency, rate or extent of a biological process.
INACTIVATION	Process that results in the transition of an entity from an active to an inactive state.

Please refer to the corresponding types in SBO and GO for the full definitions and the annotation guidelines for details on their annotation. For discussion of the relation between these types and other representations applied in pathway models, we refer to [15].

5.3 Event argument roles

The role in which each event argument (entity/other event) participates in an event is specified as one of the following.

Theme the entity/event that undergoes the effects of the event. For example, the entity that is transcribed in a TRANSCRIPTION event or transported in a TRANSPORT event.

Cause the entity/event causing the event. Marks, for example, "P1" in "P1 inhibits P2 expression". *Cause* can only be specified for events of type REGULATION and its subtypes.

At,From,To : the location in which the *Theme* entity of a LOCALIZATION event is localized (*At*) in LOCALIZATION events not involving movement or is transported (or moves) from/to (*From*/*To*) in LOCALIZATION and TRANSPORT events involving movement.

Site the site on another participating entity that is modified in the event. Can be specified for modification events such as PHOSPHORYLATION.

Participant General role type identifying an entity that participates in some underspecified way in a high-level process. Only applied for participants in annotations of the PATHWAY type.

5.4 Secondary annotation targets

In addition to the primary categories of event and entity annotation, we mark two secondary types of annotation: equivalence relations and two event modifications, NEGATION and SPECULATION. Equivalence relations are represented as undirected binary relations and event modifications as binary flags modifying events. Both are in terms of annotation scope and sematics identically to their definitions in the BioNLP ST'09 [10].

5.5 Annotation process

The event annotation assumes the entity annotation (Section 4) as its input, and the event annotation task is thus to identify references to events of the annotated types in text, their participants, and to create event structures (Figure 2).

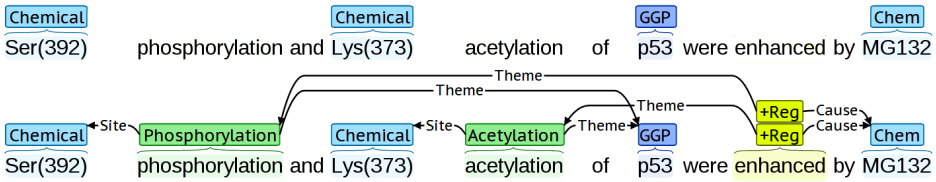

Figure 2: Illustrations of event annotation. Above: input with entity annotations only. Below: completed event annotation.

To assure that the quality and consistency of the event annotation are as high as possible, the event annotation will be created (on the basis of the entity annotation) entirely manually, without automatic support. This annotation effort will be carried out using the BRAT annotation tool [18] by a group of biologists in collaboration between NaCTeM and KISTI.

References

[1] M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. *Nature genetics*, 25:25–29, 2000.

[2] The UniProt Consortium. Ongoing and future developments at the universal protein resource. *Nucleic Acids Research*, 39(suppl 1):D214–D219, 2011.

[3] Peter Corbett and Ann Copestake. Cascaded classifiers for confidence-based chemical named entity recognition. *BMC Bioinformatics*, 9(Suppl 11):S4, 2008.

[4] M. Courtot, N. Juty, C. Knüpfer, D. Waltemath, A. Zhukova, A. Dräger, M. Dumontier, A. Finney, M. Golebiewski, J. Hastings, et al. Controlled vocabularies and semantics in systems biology. *Molecular systems biology*, 7(1), 2011.

[5] K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. Mcnaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(suppl 1):D344–D350, 2008.

[6] Robert D. Finn, Jaina Mistry, John Tate, Penny Coggill, Andreas Heger, Joanne E. Pollington, O. Luke Gavin, Prasad Gunasekaran, Goran Ceric, Kristoffer Forslund, Liisa Holm, Erik L. L. Sonnhammer, Sean R. Eddy, and Alex Bateman. The pfam protein families database. *Nucleic Acids Research*, 38(suppl 1):D211–D222, 2010.

[7] D.M. Jessop, S.E. Adams, E.L. Willighagen, L. Hawizy, and P. Murray-Rust. Oscar4: a flexible architecture for chemical text-mining. *Journal of cheminformatics*, 3(1):1–12, 2011.

[8] B. Kemper, T. Matsuzaki, Y. Matsuoka, Y. Tsuruoka, H. Kitano, S. Ananiadou, and J. Tsujii. Pathtext: a text mining integrator for biological pathway visualizations. *Bioinformatics*, 26(12):i374–i381, 2010.

[9] J.D. Kim, N. Nguyen, Y. Wang, J. Tsujii, T. Takagi, and A. Yonezawa. The genia event and protein coreference tasks of the bionlp shared task 2011. *BMC Bioinformatics*, 13(Suppl 11):S1, 2012.

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

[11] N. Le Novere, B. Bornstein, A. Broicher, M. Courtot, M. Donizelli, H. Dharuri, L. Li, H. Sauro, M. Schilstra, B. Shapiro, et al. Biomodels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic acids research*, 34(suppl 1):D689–D691, 2006.

[12] Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic Acids Research*, 33(suppl 1):D54, 2005.

[13] H. Mi, N. Guo, A. Kejariwal, and P.D. Thomas. Panther version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic acids research*, 35(suppl 1):D247–D252, 2007.

[14] T. Ohta, S. Pyysalo, J. Tsujii, and S. Ananiadou. Open-domain anatomical entity mention detection. In *Association for Computational Linguistics*, page 27, 2012.

[15] Tomoko Ohta, Sampo Pyysalo, and Jun'ichi Tsujii. From pathways to biomolecular events: Opportunities and challenges. In *Proceedings of BioNLP 2011 Workshop*. Association for Computational Linguistics, 2011.

[16] S. Pyysalo, T. Ohta, R. Rak, D. Sullivan, C. Mao, C. Wang, B. Sobral, J. Tsujii, and S. Ananiadou. Overview of the id, epi and rel tasks of bionlp shared task 2011. *BMC bioinformatics*, 13(Suppl 11):S2, 2012.

[17] R. Rak, A. Rowley, W. Black, and S. Ananiadou. Argo: an integrative, interactive, text mining-based workbench supporting curation. *Database: The Journal of Biological Databases and Curation*, 2012, 2012.

[18] P. Stenetorp, S. Pyysalo, G. Topic, T. Ohta, S. Ananiadou, and J. Tsujii. Brat: a web-based tool for nlp-assisted text annotation. *EACL 2012*, page 102, 2012.

[19] John Wilbur, Lawrence Smith, and Lorraine Tanabe. BioCreative 2. Gene Mention Task. In L. Hirschman, M. Krallinger, and A. Valencia, editors, *Proceedings of Second BioCreative Challenge Evaluation Workshop*, pages 7–16, 2007.