Cancer Genetics (CG) task at BioNLP Shared Task 2013 Task Proposal

Contents

1 Introduction **3**

2 Representation and Task Setting **3**

3 Document Selection **4**

4 Entity Annotation **5**

4.1 Annotation overview 5

4.2 Annotation process 7

5 Event annotation **8**

5.1 Annotation overview 8

5.2 Event argument roles 9

5.3 Annotation process 9

1 Introduction

This document presents the Cancer Genetics (CG) task proposed for the BioNLP Shared Task 2013.

The CG task aims to advance the automatic extraction of information from statements on the biological processes relating to the development and progression of cancer.

The BioNLP Shared Task series has been instrumental in encouraging the development of methods and resources for the automatic extraction of bio-processes from text, but efforts within this framework have been almost exclusively focused on molecular and sub-cellular level entities and events. For many domains of life sciences, entities and events (or processes) at higher levels of biological organization such as cell proliferation, apoptosis, blood vessel development, and organ growth are equally important. In particular, to be relevant to cancer biology, event extraction technology should be generalized to be able to extract statements relating molecular-level entities and events to anatomy- level effects and organism-level outcomes. We aim to advance the development of such event extraction methods and the capacity of automatic analysis of texts on cancer biology through the Cancer Genetics (CG) task of BioNLP Shared Task 2013.

2 Representation and Task Setting

The BioNLP Shared Task (ST) 2013 Cancer Genetics (CG) 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. CELLULAR COMPONENT, CELL, TISSUE, ORGAN). 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. DEVELOPMENT, DEATH), 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 CG 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 the main tasks in ST'09 and ST'11 and will thus be familiar to participants who have taken part in

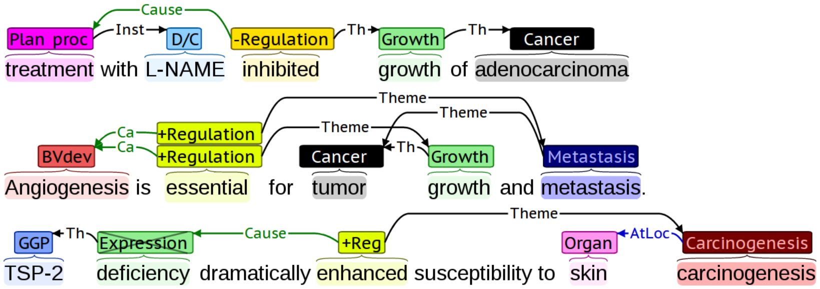
Cause
Inst
Th
Th
Plan proc
D/C
-Regulation
Growth
Cancer
treatment with L-NAME inhibited growth of adenocarcinoma
Theme
Theme
Theme
Ca
Ca
+Regulation
+Regulation
Th
BVdev
Cancer
Growth
Metastasis
Angiogenesis is essential for tumor growth and metastasis.
Theme
Th
Cause
AtLoc
GGP
Expression
+Reg
Organ
Carcinogenesis
TSP-2 deficiency dramatically enhanced susceptibility to skin carcinogenesis

Figure 1: Illustrations of event annotations.

these events and compatible with existing systems introduced to address existing tasks. Additionally, this task setting allows the application of the evaluation criteria 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 as a random sample of documents relevant to a selection of subdomains central to cancer biology [8]:

- Evasion of programmed cell death, **apoptosis**
- Blood supply through new blood vessel development, **angiogenesis**
- Invasion and **metastasis**
- Modification of **growth signaling**
- Energy supply, **glucose metabolism**

And the **mutations** leading to aberrations in these processes. Specifically, the documents are selected through PubMed queries, iteratively refined to assure document relevance, optionally combined with manual relevance filtering when necessary.

For one of these subdomains, angiogenesis, the CG task will make use of a recently introduced corpus, MLEE [12]. The event annotation of this corpus has been recast and extended to match the CG task annotation guidelines.

4 Entity Annotation

This section introduces the entity annotation applied in the CG 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 defined for the CG task (labels in gray are categories defined only for organization and are not used in evaluation):

Type	Scope	Reference
ORGANISM	organism mentions	taxonomy DBs
ANATOMICAL ENTITY	structural organization of organism	CARO
MATERIAL ANATOMICAL ENTITY	anatomical entities w/mass	CARO
ANATOMICAL STRUCTURE	material ASs with structure	CARO
ORGANISM SUBDIVISION	fiat parts of multicellular organism	CARO
ANATOMICAL SYSTEM	ASs of multiple organs	CARO
ORGAN	ASs of multiple multi-tissue structs.	CARO
MULTI-TISSUE STRUCTURE	AS of multiple tissues	CARO
TISSUE	ASs of similar cells and ECM	CARO
DEVELOPING STRUCTURE	ASs varying in granularity due to development	CARO
CELL	ASs of cell compartment, surrounded by PM	CL
CELLULAR COMPONENT	ASs that are parts of cells	GO-CC
ORGANISM SUBSTANCE	gaseous, liquid or semisolid material ASs	CARO
IMMATERIAL ANATOMICAL ENTITY	anatomical entities without mass	CARO
MOLECULAR ENTITY		
GENE OR GENE PRODUCT	genes, RNA and proteins	gene/protein DBs
CHEMICAL	simple, non-repetitive chemical entities	ChEBI
PROTEIN DOMAIN OR REGION	parts of protein molecules	-
DNA DOMAIN OR REGION	short, specifically identified spans of DNA	-
PATHOLOGICAL FORMATION	pathological material organism parts	-
CANCER	cancerous pathological formations	-

GENE OR GENE PRODUCT and 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. 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 [3], Entrez Gene [10] and PFam [6].

CELLULAR COMPONENT is annotated following the scope of the Gene Ontology `cellular component` subontology [1]. CELL is annotated following the scope of the Cell Ontology (CL) [2], expanded to incorporate also pathological variants of cells that are part of canonical (healthy) anatomy such as cancer cells.

Mentions of anatomical structures at leves of granularity between the cell and the whole organism as well as organism substances (e.g. blood) and immaterial anatomical entities (e.g. lumen) are annotated following the Common Anatomy Reference Ontology (CARO) [7], using compatible species-specific OBO resources as reference. The scope of these entity annotations is likewise expanded to cover also pathological variants (e.g. cancerous liver).

Finally, pathological material organism parts that have no correspondence in canonical anatomy (e.g. scars) are annotated as PATHOLOGICAL FORMATION, with those that refer to cancerous entities marked more specifically as CANCER.

In terms of mention types in text, the annotation for CHEMICAL and GENE OR GENE PRODUCT covers entity name mentions only, while the annotation for other entity types covers entity name mentions, nominal mentions, and adjectival mentions (e.g. "*cancerous*").

Further details are given in the task annotation guidelines.

4.2 Annotation process

The base entity annotation for the CG corpus will be created automatically using state-of-the-art entity mention taggers for each of the targeted entity types.

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
CHEMICAL	OSCAR4
GENE OR GENE PRODUCT	NERsuite
Anatomical and pathological entities	NERsuite

For CHEMICAL tagging, the OSCAR4 system [9] trained on the chemical named entity recognition corpus of [4] will be applied. For GENE OR GENE PRODUCT mention detection, the NERsuite[1] system trained on the BioCreative 2 Gene Mention task [14] corpus will be used. NERsuite will also be applied for the detection of anatomical entities and pathological formation detection, for this task trained on an updated combination of the Multi-Level Event Extraction (MLEE) [12] and Anatomical Entity Mention (AnEM) [11] corpora.

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, a level of performance 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 for this task.

5 Event annotation

This section introduces the event annotation applied in the CG 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 CG task:

Event type	Arguments
Anatomical	
Development	*Theme*(Anatomical/Pathological)
Blood vessel development	*Theme*(Anatomical/Pathological), *At*(Anatomical)
Growth	*Theme*(Anatomical/Pathological)
Death	*Theme*(Anatomical/Pathological)
Apoptosis	*Theme*(Cell)
Breakdown	*Theme*(Anatomical/Pathological)
Cell proliferation	*Theme*(Cell)
Cell division	*Theme*(Cell)
Remodeling	*Theme*(Tissue)
Reproduction	*Theme*(Organism)
Pathological	
Mutation	*Theme*(GGP)
Carcinogenesis	*Theme*(Anatomical/Pathological), *At*(Anatomical)
Metastasis	*Theme*(Anatomical/Pathological), *To*(Anatomical)
Molecular	
Metabolism	*Theme*(Molecule)
Synthesis	*Theme*(Chemical)
Catabolism	*Theme*(Molecule)
Gene expression	*Theme*(GGP)
Transcription	*Theme*(GGP)
Translation	*Theme*(GGP)
Protein processing	*Theme*(GGP)
Phosphorylation	*Theme*(GGP), *Site*(PDR)
Dephosphorylation	*Theme*(GGP), *Site*(PDR)
Acetylation, Ubiquitination, etc. (defined similarly to Phosphorylation)	
DNA methylation	*Theme*(GGP), *Site*(DDR)
DNA demethylation	*Theme*(GGP), *Site*(DDR)
Pathway	*Participant*(Molecule)+
General	
Localization	*Theme*(Entity), *At/To/From*(Entity)
Binding	*Theme*(Entity)+, *Site*(PDR or DDR)+
Dissociation	*Theme*(Entity)+, *Site*(PDR or DDR)+
Regulation	*Theme*(ANY), *Cause*(ANY), *Site*(PDR or DDR)
Positive regulation	*Theme*(ANY), *Cause*(ANY) , *Site*(PDR or DDR)
Negative regulation	*Theme*(ANY), *Cause*(ANY), *Site*(PDR or DDR)
Planned process	*Theme*(ANY), *Instrument*(Entity)

Here, "Molecule" refers to an entity annotation of the type Chemical or Gene or gene product, "Anatomical" ("Pathological") to one of any of the anatomical (pathological) entity types (Section 4), "Entity" to any entity type, and "ANY" 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 Gene Ontology (GO) [1].

5.2 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 develops in a DEVELOPMENT 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.

Instrument entity used to carry out a planned process.

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.

5.3 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.

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 [13]. The annotation effort will be coordinated

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

by Tomoko Ohta, and the setup and use of the annotation software and tools for the evaluation of inter-annotator agreement will be supported by Sampo Pyysalo.

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] J. Bard, S.Y. Rhee, and M. Ashburner. An ontology for cell types. *Genome biology*, 6(2):R21, 2005.

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

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

[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] Melissa A. Haendel, Fabian Neuhaus, David Osumi-Sutherland, Paula M. Mabee, José L.V. Mejino, Chris J. Mungall, and Barry Smith. CARO–the common anatomy reference ontology. *Anatomy Ontologies for Bioinformatics*, pages 327–349, 2008.

[8] D. Hanahan and R.A. Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.

[9] 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.

[10] 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.

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

[12] S. Pyysalo, T. Ohta, M. Miwa, H.C. Cho, J. Tsujii, and S. Ananiadou. Event extraction across multiple levels of biological organization. *Bioinformatics*, 28(18):i575–i581, 2012.

[13] 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.

[14] 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.