BioID Track Overview*
BioCreative VI
Bethesda, MD
Oct. 18, 2017

Cecilia Arighi¹, Lynette Hirschman², Thomas Lemberger³, Samuel Bayer², Robin Liechti⁴, Donald Comeau⁵, Cathy Wu¹

¹Dept. of Computer and Information Sciences, U of Delaware, Newark, DE, USA
²Information Technology Center, The MITRE Corporation, Bedford, M, USA
³EMBO, Heidelberg, Germany
⁴SIB, Lausanne, Switzerland
⁵NCBI, NIH, Bethesda, MD, USA

*This work was supported under NIH grant 5R01GM080646-11S1 and NIH R13 GM109648
BioID Track

- **Vision**
  - Facilitate article curation & indexing at pre- & post-publication stages

- **Data**
  - Figure panel captions, extracted and annotated by SourceData
  - Supports capture of major experimental findings in terms of relations among bioentity IDs
  - Makes findings searchable, linked across papers

- **Tasks**
  - Bioentity Tagging
    - Entity tagging: annotating all mentions of key bioentity classes, e.g., protein, small molecules,…
    - Normalization: ID assignment to link bioentity mentions to identifiers in biological resources, e.g., UniProt, ChEBI,…
  - Interactive tools: Assist humans (authors, curators) in annotation
Batch Processing (Part 1)

U Delaware
- Organized track
- Released training, test sets
- Created equivalence classes for normalization

SourceData
- Extracted figure panel captions
- Provided annotations in XML

MITRE
- Converted SourceData annotations to BioC
- Developed scorer & scored results

NCBI
- Provided full text articles in BioC

Participants
- Provided annotations for bioentities
BioID Track: Pipeline for Interactive Experiment

Interactive Processing (Part 2)

U Delaware
- Recruiting curators, authors
- Developing framework for interactive experiment

MITRE
- Converting BioC to PubAnnotation for ingest into SourceData

SourceData
- Ingesting annotations from high scoring system into SourceData annotation tool

Curators, Authors
- Performing annotations w/wo BioID track pre-annotations

SourceData, U Delaware
- Comparing annotation quality, efficiency of annotation process
(c) Cells were treated with 200nM epoxomicin for 4h and then immunoblotted and probed with anti-pTyr-416-Src, anti-Src, anti-p53 and anti-actin antibodies. Uncropped images of blots are shown in Supplementary Fig. S9.
### BioEntities and Associated Resources (SourceData Training Set)

<table>
<thead>
<tr>
<th>BioEntity Class</th>
<th>Resources for ID (# articles w entity)</th>
<th>Total # by entity type</th>
<th># unique by entity type</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Protein/Gene (561 articles)</strong></td>
<td>UniProt (548)</td>
<td>30,211</td>
<td>2,833</td>
</tr>
<tr>
<td></td>
<td>NCBI gene (537)</td>
<td>21,766</td>
<td>2,451</td>
</tr>
<tr>
<td><strong>miRNA</strong></td>
<td>9</td>
<td>167</td>
<td>13</td>
</tr>
<tr>
<td><strong>Small molecules (513 articles)</strong></td>
<td>ChEBI (506)</td>
<td>9,869</td>
<td>786</td>
</tr>
<tr>
<td></td>
<td>PubChem (134)</td>
<td>700</td>
<td>140</td>
</tr>
<tr>
<td><strong>Cellular component</strong></td>
<td>GO (482)</td>
<td>7,310</td>
<td>376</td>
</tr>
<tr>
<td><strong>Cell types &amp; cell lines (482 articles)</strong></td>
<td>Cellosaurus (351)</td>
<td>5,783</td>
<td>230</td>
</tr>
<tr>
<td></td>
<td>Cell Ontology (300)</td>
<td>4,638</td>
<td>217</td>
</tr>
<tr>
<td><strong>Tissues &amp; organs</strong></td>
<td>Uberon (316)</td>
<td>5,870</td>
<td>459</td>
</tr>
<tr>
<td><strong>Organisms &amp; species</strong></td>
<td>NCBI Taxonomy (454)</td>
<td>7,952</td>
<td>147</td>
</tr>
</tbody>
</table>
Annotation Guidelines

- **Entities are assigned to only one entity type**
  - All mentions of a bioentity are tagged with an identifier from the appropriate resource
  - Resources are listed in preference order, e.g., for small molecules, use ChEBI (first choice) otherwise PubChem
  - Certain terms may be typed but not normalized, e.g., “GFP” is tagged as a protein, but not normalized

- **Generic terms referring to broad classes of biological components (e.g. 'proteins', 'cells’) are not be tagged**
  - Unless they refer to the object of an assay

- **Only the base term is tagged, not prefixes or suffixes**
  - In ‘B-RAF(V600E)’ only B-RAF is tagged
  - Except for prefixes indicating species, e.g., ‘dMyc’ is tagged because the prefix d denotes the homolog in Drosophila of Myc
A. The localization of NSUN3 was analysed in HEK293 cells stably expressing NSUN3-GFP (green). NSUN3-GFP and staining with a Mitotracker (red) are shown separately and in an overlay with DAPI to indicate nuclei. The scale bar represents 5 m.

Annotation of NSUN3, linked to UniProt ID Q9H649

Annotation of GFP as protein (not linked)
Scoring

- Participant annotations are scored by comparing against the reference (SourceData) annotations
  - Score based on standard measures of precision, recall, F1-measure

- Mention level scoring
  - Scores are generated for “Normalized mentions” & “Any mentions”
  - Matches are either “exact” or “overlap”

<table>
<thead>
<tr>
<th>Mention Scoring</th>
<th>Any Mention</th>
<th>Normalized Mentions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact span match</td>
<td>Any Exact</td>
<td>Norm Exact</td>
</tr>
<tr>
<td>Span Overlap</td>
<td>Any Overlap</td>
<td>Norm Overlap</td>
</tr>
</tbody>
</table>

- Overlap scores ≥ Exact scores (by definition)
- In general, “Any Mention” scores are higher than “Normalized Mention” scores

- Normalization
  - Score is computed against the set of unique identifiers in a caption
## Results Across BioEntities

<table>
<thead>
<tr>
<th>Entity Type</th>
<th>Mention - All Overlap</th>
<th>Normalization Micro Avg</th>
<th># Teams</th>
<th># Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Team</td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>Cell type or line</td>
<td>407</td>
<td>0.84</td>
<td>0.76</td>
<td><strong>0.80</strong></td>
</tr>
<tr>
<td>Cellular component</td>
<td>407</td>
<td>0.73</td>
<td>0.55</td>
<td><strong>0.63</strong></td>
</tr>
<tr>
<td>Gene or protein</td>
<td>407</td>
<td>0.83</td>
<td>0.84</td>
<td><strong>0.83</strong></td>
</tr>
<tr>
<td>Organism or species</td>
<td>407</td>
<td>0.88</td>
<td>0.83</td>
<td><strong>0.85</strong></td>
</tr>
<tr>
<td>Small molecule</td>
<td>407</td>
<td>0.80</td>
<td>0.60</td>
<td><strong>0.69</strong></td>
</tr>
<tr>
<td>Tissue or organ</td>
<td>407</td>
<td>0.79</td>
<td>0.63</td>
<td><strong>0.70</strong></td>
</tr>
<tr>
<td>miRNA</td>
<td>386</td>
<td>not scored</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3 entities have mention overlap > 0.8 (Cell type; Gene/protein; Species)

Highest normalization score: Species (5 teams, 9 runs)
Top Mention Scores (Any Overlap)

F1-scores for cell type, gene/protein and organism are greater or equal to 0.80
Top Normalized Scores (Micro-Average)

Normalized scores are lower:
All F1-scores < 0.8
Species and Gene/Protein Results

Species mention scores and normalized scores are close (silver, gold stars)

Gene/protein mention scores (blue/green triangles) are higher than normalization scores (red triangles)
BioID Track Deliverables

- **Large training and test data sets**
  - Annotated with 6 bioentity classes
  - Normalized to standard database identifiers

- **Conversion tools for BioC ⇔ PubAnnotation format**
  - Supporting release of training and test sets in BioC
  - Ingest of participant annotations back into SourceData format

- **Scorer**
  - For mention level scoring
  - For normalization
    - Including support for equivalence classes of identifiers: UniProt/EntrezGene and ChEBI/PubChem
Next Steps

- **Complete interactive curation experiment**
  - Have curators, authors use pre-annotations from the BioID track

- **Do error analysis**
  - Why are mention level scores for “normalized” subset lower than “any mention” scores?
  - Why were scores low for gene/protein normalization?
    - Failure to identify organism correctly?
    - Need to analyze complex expressions containing gene/protein names?
  - How do these results compare to state of the art?