BC4GO Corpus – training set description

Full text articles

The BC4GO training corpus includes 100 full-text articles from the PubMed Central (PMC) open access subset. These 100 articles are made available in BioC XML format (see pmc_go.key for details) with the goal of providing easier access to the text content than that provided by the traditional PMC XML data model, which is designed to preserve all original article details without loss and thus incorporates great flexibility to meet the organization and display needs of many different publishers. To access all of the detailed information in an article, PMC XML is the format of choice. If one wants to access the text of an article, then BioC is a more convenient choice. Each paper is named using its PubMed identifier as pmid.xml (e.g. 20130316.xml).

Annotation files

BC4GO training set also contains 2,933 annotations for those 100 articles, made by expert GO curators from five Model Organism Databases (MODs) (See details in Table below).

MOD	FlyBase	MaizeGDB	RGD	TAIR	WormBase
Annotations	647	517	687	349	733

Each annotation file is linked to one paper by sharing the same PMID in its file name and includes all annotations of the paper. Each annotation has its own unique ID and is defined by four distinct elements:

GO evidence sentence: we provide both the text and its character offsets in the paper. Note that an annotation may be linked to more than one evidence sentence. If multiple sentences are located in the same passage, those sentences are listed sequentially (e.g. annotation id=23593298_52 in annotation_23593298.xml). Otherwise, for annotations with sentences in multiple passages, sentences are shown in separate passages but can be linked using the same annotation id (e.g. two occurrences of annotation id=12213836_44 in annotation_12213836_44.xml).

GO term: we provide both the term itself and its GO id, separated by pipes.

GO evidence code[1]: we provide its abbreviated name (e.g. IDA for Inferred from Direct Assay). When the evidence sentence is of Summary type and evidence code information is unavailable, NONE is used instead.

[1] http://www.geneontology.org/GO.evidence.shtml

Gene: we provide both its name and corresponding NCBI Gene ID when available (separated by pipes). Note that, one GOES may be used for making annotations for multiple genes in the same sentence. In such cases, separate annotations will be made, one for each gene.

Below we show one example of such an annotation in the BioC XML format.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE collection SYSTEM "BioC.dtd">
<collection>
        <source>GO_Annotation</source>
        <date>20130316</date>
        <key>go_annotation.key</key>
        <document>
                <id>16987298</id>
                <passage>
                        <infon key="type">paragraph</infon>
                        <offset>21171</offset>
                        <annotation id="16987298_1">
                                <infon key="gene">Fas(246097)</infon>
                                <infon key="go-term">response to organic cyclic
                                compound|GO:0014070</infon>
                                <infon key="goevidence">IEP</infon>
                                <infon key="type">GOA</infon>
                                <location length="171" offset="21273"/>
                                <text>Results showed that BP significantly induced Fas expression
                        (from one to 41.3-fold vs. one to 7.7-fold in DBTRG-05MG and RG2 cells,
                        respectively) but not Fas-L expression.</text>
                        </annotation>
                </passage>
                ...
```

Acknowledgment

We would like to thank Drs. Don Comeau, Rezarta Dogan and John Wilbur for their technical assistance of the BioC XML format.