Toward Community-Driven, Shared Literature Annotation Resources

Jin-Dong Kim
Database Center for Life Science (DBCLS)
DBCLS

- Database Center for Life Science
  - A government-funded research center
  - For integration of databases of life sciences
  - It annually organizes
    - BioHackathon series
    - BLAH series
DBCLS

Research Organization

- ROIS
  - Informatics
    - ...
    - NII
  - Genetics
    - NIG
    - DDBJ
- DBCLS

Funding Agency

- JST
- NBDC

Service

Database

R&D
Introduction to DBCLS

Research Organization

ROIS

Informatics

... NII

Genetics

NIG

DBCLS

Database

R&D

Funding Agency

JST

NBDC

Service

PDBj

DDBJ
Introduction to DBCLS
“Inheritance involves the passing of discrete units of inheritance”

Proceedings of the Natural History Society of Brünn, 1866

Gregor Mendel
“Inheritance involves the passing of discrete units of inheritance”

Proceedings of the Natural History Society of Brünn, 1866

Gregor Mendel
Heredity

geneome
“all media are extensions of our human senses, bodies and minds.”

*The Medium Is the Massage*, 1967
Extension of Heredity

genome

literature
Cell

DNA

mRNA Transcription

Mature mRNA

Nucleus

Transport to cytoplasm for protein synthesis (translation)

tRNA

mRNA

Cell membrane
Society

transcribed

translated

transcribed
Science Literature

- Scholarly articles, Textbooks, ...
- Authoritative information
- Accumulation of scientific knowledge
- Basis for new discoveries
- Repeatedly accessed
My research is based on ...
Literature Indexing
Geospatial Indexing
Google Maps (Washington D.C.)
(Geospatial) Pathways
Entities

ROI production and NF-kappaB activation, which could both be blocked by antioxidants or FLAP inhibitors
(Linguistic) Pathways

ROI production and NF-kappaB activation, which could both be blocked by antioxidants or FLAP inhibitors.
Why people use Google Maps?

- Useful
  ✓ Contents

- Easy to use
  ✓ Interface
    ➔ To access
    ➔ To exchange
    ➔ To create
    ➔ To reuse

Geospatial annotations
Literature annotation

• Do we have good contents?
  ✔ Many groups are producing annotations.

• Do we have good ways to access them?
  ✔ Ann. resources are scattered and isolated.

Let’s link them to each other, and Share them, altogether.  **BLAH!**
PubAnnotation

✔ Is a repository of literature annotation
✔ Is based on a scalable storage system
✔ Specifically aims at PubMed and PMC
✔ Solicits contribution of annotations from the community
✔ Solves various problems for sharing the annotations
  ➔ Alignment
  ➔ Global addressing system
  ➔ REST APIs
Challenges for Integration

● Format is not standardized
  ✔ Many proprietary formats

● Texts are changed
  ➔ PubMed, PMC change texts
  ➔ Web masters change texts
  ➔ Annotation projects change texts

✔ For
  ➔ Cleaning
  ➔ Convenience for annotation
    – Unicode → ASCII
Challenges for Integration

- Format is not standardized
  - A matter of conversion
- Texts are changed
  - Breaks stand-off annotation
    - Character offsets become invalid
  - Solution
    - Sequence alignment (BLAST!)
GATA3-Driven Th2 Responses Inhibit TGF-1Induced FOXP3 Expression and the Formation of Regulatory T Cells Transcription factors act in concert to induce lineage commitment towards Th1, Th2, or T regulatory (Treg) cells, and their counter-regulatory mechanisms were shown to be critical for polarization between Th1 and Th2 phenotypes. FOXP3 is an essential transcription factor for natural, thymus-derived (nTreg) and inducible Treg (iTreg) commitment; however, the mechanisms regulating its expression are as yet unknown. We describe a mechanism controlling iTreg polarization, which is overruled by the Th2 differentiation pathway. We demonstrated that interleukin 4 (IL-4) present at the time of T cell priming inhibits FOXP3. This inhibitory mechanism was also confirmed in Th2 cells and in T cells of transgenic mice overexpressing GATA-3 in T cells, which are shown to be deficient in transforming growth factor (TGF) β-mediated FOXP3 induction. This inhibition is mediated by direct binding of GATA3 to the FOXP3 promoter, which represses its transactivation process. ...
GATA3-Driven Th2 Responses Inhibit TGF-1Induced FOXP3 Expression and the Formation of Regulatory T Cells Transcription factors act in concert to induce lineage commitment towards Th1, Th2, or T regulatory (Treg) cells, and their counter-regulatory mechanisms were shown to be critical for polarization between Th1 and Th2 phenotypes. FOXP3 is an essential transcription factor for natural, thymus-derived (nTreg) and inducible Treg (iTreg) commitment; however, the mechanisms regulating its expression are as yet unknown. We describe a mechanism controlling iTreg polarization, which is overruled by the Th2 differentiation pathway. We demonstrated that interleukin 4 (IL-4) present at the time of T cell priming inhibits FOXP3. This inhibitory mechanism was also confirmed in Th2 cells and in T cells of transgenic mice overexpressing GATA-3 in T cells, which are shown to be deficient in transforming growth factor (TGF) β-mediated FOXP3 induction. This inhibition is mediated by direct binding of GATA3 to the FOXP3 promoter, which represses its transactivation process. 

This inhibitory mechanism was also confirmed in Th2 cells of transgenic mice over-expressing GATA-3 in T cells, which are shown to be deficient in transforming growth factor (TGF)-beta-mediated FOXP3 induction.
GATA3-Driven Th2 Responses Inhibit TGF-1Induced FOXP3 Expression and the Formation of Regulatory T Cells

Transcription factors act in concert to induce lineage commitment towards Th1, Th2, or T regulatory (Treg) cells, and their counter-regulatory mechanisms were shown to be critical for polarization between Th1 and Th2 phenotypes. FOXP3 is an essential transcription factor for natural, thymus-derived (nTreg) and inducible Treg (iTreg) commitment; however, the mechanisms regulating its expression are as yet unknown. We describe a mechanism controlling iTreg polarization, which is overruled by the Th2 differentiation pathway. We demonstrated that interleukin 4 (IL-4) present at the time of T cell priming inhibits FOXP3. This inhibitory mechanism was also confirmed in Th2 cells and in T cells of transgenic mice overexpressing GATA-3 in T cells, which are shown to be deficient in transforming growth factor (TGF-β)-mediated FOXP3 induction. This inhibition is mediated by direct binding of GATA3 to the FOXP3 promoter, which represses its transactivation process. ...
Alignment


✔ A definite solution to text variant problem
✔ It can align even full paper articles sourced by two different groups.
Aligned annotations from different groups
Alignment

• Aligned annotations

✓ http://www.pubannotation.org/
Global addressing system

- Persistently preserve the texts of all the articles from PubMed / PMC(OA)
  - UTF-8
  - ASCII conversion is provided
- Offset indices are stably maintained
a case of Google Map
A case of PubAnnotation

● Example of URL
  ✔ http://pubannotation.org/docs/sourcedb/PubMed/sourceid/10022882/spans/606-710/annotations/visualize

● How to get the URL (Example)
  ✔ http://pubannotation.org/docs/sourcedb/PubMed/sourceid/10022882
Global addressing system

• Persistently preserve the texts of all the articles from PubMed / PMC(OA)
  ✓ UTF-8
  ✓ ASCII conversion is provided

• Offset indices are stably maintained
Hi Bill, what is the diagnostic test for MERS infection?

Check this link out. FYR, I've annotated it using NCIt, OBI, and SNOMEDCT.
Hi Bill, what is the diagnostic test for MERS infection?

Check this link out. FYR, I've annotated it using NCIt, OBI, and SNOMEDCT.

Currently, in PubAnnotation

### Documents

<table>
<thead>
<tr>
<th>Name</th>
<th>PMC</th>
<th>PubMed</th>
<th>Keywords</th>
<th>Source ID</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMC</td>
<td>9,044</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PubMed</td>
<td>8,424,657</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FirstAuthor</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Projects (146)

### With most annotations

<table>
<thead>
<tr>
<th>Name</th>
<th># Ann.</th>
<th>Updated At</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubmedHPO</td>
<td>12,437,742</td>
<td>2016-01-31</td>
<td>Uploading</td>
</tr>
<tr>
<td>DisGeNET</td>
<td>3,117,504</td>
<td>2016-01-28</td>
<td>Beta</td>
</tr>
<tr>
<td>NEUROSES</td>
<td>2,151,082</td>
<td>2016-02-24</td>
<td>Beta</td>
</tr>
<tr>
<td>CRAFT-treebank</td>
<td>844,123</td>
<td>2015-11-19</td>
<td>Beta</td>
</tr>
<tr>
<td>bionlp-st-seedev-2016-training</td>
<td>225,680</td>
<td>2016-05-25</td>
<td>Released</td>
</tr>
<tr>
<td>spacy-test</td>
<td>136,597</td>
<td>2016-05-25</td>
<td>Released</td>
</tr>
<tr>
<td>FSU-PRGE</td>
<td>59,505</td>
<td>2016-05-17</td>
<td>Released</td>
</tr>
<tr>
<td>craft</td>
<td>52,960</td>
<td>2015-10-13</td>
<td>Beta</td>
</tr>
<tr>
<td>jnlpba-st-training</td>
<td>51,290</td>
<td>2016-09-09</td>
<td>Released</td>
</tr>
<tr>
<td>PennBioIE</td>
<td>23,881</td>
<td>2016-05-17</td>
<td>Released</td>
</tr>
</tbody>
</table>

### Recently updated

<table>
<thead>
<tr>
<th>Name</th>
<th># Ann.</th>
<th>Updated At</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>jnlpba-st-training</td>
<td>51,290</td>
<td>2016-09-09</td>
<td>Released</td>
</tr>
<tr>
<td>jnlpba-st-test</td>
<td>6,005</td>
<td>2016-09-07</td>
<td>Uploading</td>
</tr>
<tr>
<td>bionlp-st-seedev-2016-training</td>
<td>109</td>
<td>2016-09-05</td>
<td>Uploading</td>
</tr>
<tr>
<td>bionlp-st-bb3-2016-training</td>
<td>1,292</td>
<td>2016-09-05</td>
<td>Released</td>
</tr>
<tr>
<td>bionlp-st-cg-2013-training</td>
<td>10,935</td>
<td>2016-08-22</td>
<td>Released</td>
</tr>
<tr>
<td>bionlp-st-epi-2011-training</td>
<td>7,595</td>
<td>2016-08-22</td>
<td>Released</td>
</tr>
<tr>
<td>Ab3P-abbreviations</td>
<td>2,342</td>
<td>2016-07-29</td>
<td>Beta</td>
</tr>
<tr>
<td>AnEM_full-texts</td>
<td>689</td>
<td>2016-07-27</td>
<td>Uploading</td>
</tr>
</tbody>
</table>

### News

- (09 Mar 2016) Recent access problem
- (25 Feb 2016) System problem on 25/02/2016
- (19 Jan 2016) a new project status, Uploading, added.
- (04 Jan 2016) News service begins.
PubAnnotation

● New version to be released soon
  ✔ Performance improved
  ✔ Interface improved
  ✔ BioC conversion to be supported
    ➔ Thanks to the NCBI team
  ✔ Bug fixes
TextAE

- To access/edit annotations
  - [✓] http://textae.pubannotation.org
  - [✓] Has fully RESTful APIs
PubDictionaries

- To share dictionary resources
  - Current version
    - [http://pubdictionaries.org](http://pubdictionaries.org)
  - New version
    - [http://new.pubdictionaries.org](http://new.pubdictionaries.org)
Shared Annotation Targets

Scientific literature is a central repository of scientific knowledge - every important scientific discovery has been published in it. As such, scientific literature has become a main target of data mining, and the importance of the structured, nature of natural language texts poses a major barrier to accessing the contents of literature. Technology of literature annotation thus has played a central role for text mining. While literature annotation still requires enormous effort despite a number of years of concentrated experience, the productivity of literature annotation is recently significantly improved, and there are quite a few groups producing annotations in large scale. While many groups have released those data sets of literature annotation to the public, however, the way of sharing those widely valuable resources still remains at a primitive level, e.g., relying on individual exchange of archived files.

Meanwhile, the advancement of internet web technology has enabled much convenient ways of collaborating and producing and sharing data. For example, the technology of web 2.0 has enabled crowdsourcing and web 3.0 has enabled machine-understandable web of data. As for the annotation in general sense, for example, the Google Map system allows even laypersons to easily produce geographic annotations and immediately share them by simply adding in the URL representing the annotations. Before the era of Google Map, producing and sharing geographic annotation has never been a simple task. What made Google Map so convenient may be attributed to the sharing of the same geographic coordinates (latitude and longitude) and referencing to every geographic position by a URL, and even dereferencing of the geographic positions by URL.
Many literature annotation projects

Meanwhile, the advancement of internet web technology has enabled much convenient ways of collaborating for producing geographic annotations data. For example, the technology of web 2.0 has enabled crowdsourcing and the content generation, and web 2.0 has enabled machine-understandable web of data. In terms of annotation in general sense, for example, the Google Map system allows every user knowing how easily produce geographic annotations and immediately share them by simply appending the URL representing the annotations. Before the era of Google Map, geographic annotation was never been a simple. Google Map so convenient may be attributed to the sharing of the same coordinate system (latitude and longitude).
None of them is complete
None of them is complete
None of them is complete
Then, why don't we collect & link them?
Biomedical Linked Annotation Hackathon
BLAH

- Biomedical Linked Annotation Hackathon
  - BLAH
    - Feb. 2015, Kashiwa
  - BLAH2
    - Nov. 2015, Mishima / Ito
  - BLAHMUC
    - Oct. 2016, Munich
BLAH

• Biomedical Linked Annotation Hackathon
  ✓ BLAH
    ➔ Feb. 2015, Kashiwa
  ✓ BLAH2
    ➔ Nov. 2015, Mishima / Ito
  ✓ BLAHMUC
    ➔ Oct. 2016, Munich
  ✓ BLAH3
    ➔ Jan. 2017, Tokyo
Thank you!

Happy October Blah!