Towards Evidence Extraction: Analysis of Scientific Figures from Studies of Molecular Interactions

Gully Burns\textsuperscript{1}, Xiangyang Shi\textsuperscript{1}, Yue Wu\textsuperscript{1}, Huaigu Cao\textsuperscript{1}, Prem Natarajan\textsuperscript{1}

\textsuperscript{1} Intelligent Systems Division, USC’s Information Sciences Institute

https://sciknowengine.github.io/
Molecular Biology is both complex and beautiful

Inner Life of the Cell (https://youtu.be/wJyUtbn0O5Y)
Molecular Interactions

M1

binding site 1

? binding site 2

M2

molecular complex

M1

M2

M3

phosphorylation site

M3

phosphorylated protein
IntAct

https://www.ebi.ac.uk/intact/

- European Bioinformatics Institute @ Cambridge, England
- A large-scale, high-quality manually curated database
- Content
  - 20,065 curated papers in total
  - 2,254 open access papers
- Describes molecular interactions + detection methods linked to specific subfigures (e.g., Fig. 1b, 3d, etc.)
Scientific Cycles of Investigation

Knowledge

Questions

Data

Experiment

analyze

design

explain

execute

theory
evidence
Can we automatically classify types of scientific evidence? figures from articles?
Pipeline

Primary Experimental Literature

Figure Extraction

Cropped Figures

Subpanel Extraction

Subfigures

Captions

Subcaptions

Expt Type Classification

Expt Type + Template

Type Classification

PMC Articles (.pdf + .nxml)
1. Extracting Whole Figures

https://github.com/BMKEG/ lapdftext

1. Build a spatial index of tightly packed text blocks

2. Locate Caption Text blocks

   Regular Expression: \^Fig(ure|ure\.|)[0,1]\s*(\d+)

3. Find areas of low word density above, below and to the side of the caption

4. Remove caption text

5. Crop low word density space + caption region
indicate that the tumor progression to the most aggressive carcinoma stage in Csf1p/Csf1p PyMT mice was >10 wk delayed compared with +/Csf1p littermates (Fig. 2 D). Pun atum-assoloted Macrophages Is Correlated with Delayed Tumor Progression. CSF-1 is a macrophage growth factor, suggesting that the recruitment of macrophages to the tumor might be a factor that promotes the aggressive development of the +/Csf1p tumors. Consequently, we examined the histology of the mammary tumors and the surrounding stroma from both +/Csf1p and Csf1p/Csf1p PyMT mice at the period of the transition to carcinoma (~10 wk of age). Before the transition at 7 to 8 wk of age, a dramatic increase of leukocytic infiltration was found around the primary mammary tumors in +/Csf1p mammary glands (Fig. 3 A, indicated by arrows). No such increase was observed at the tumor site in Csf1p/Csf1p mammary glands (Fig. 3 B), though the primary tumors in both Csf1p/Csf1p and +/Csf1p PyMT mice had progressed to similar nonmalignant stages (adenoma). Immunohistochemical analysis, using a monoclonal antibody against the macrophage lineage-specific marker, F4/80, showed that a large percentage of infiltrated leukocytes in the +/Csf1p mammary gland were F4/80 positive (Fig. 3 C) and that few such cells were found around Csf1p/Csf1p primary tumors (Fig. 3 D). In concert with the histopathological development of the primary tumors to the carcinoma stage, the infiltration of leukocytes became more intense and focal infiltration sites were often seen in the mammary glands of +/Csf1p PyMT mice. Densely infiltrated cells with the morphology of granulocytes, mast cells, and monocyte-like cells were found in these sites and the tumor acini adjacent to them often displayed a disrupted boundary (Fig. 3 E, arrow). This suggests that the basement membrane of the acini at the infiltration site had lost its integrity, potentially allowing tumor cells to migrate into the adjacent connective tissue. These leukocytic infiltration sites were detected in +/Csf1p PyMT mice/alireat as 9 wk of age but were absent in Csf1p/Csf1p mammary glands at the same age (Fig. 3 F). Furthermore, an intensive infiltration of F4/80+ cells was found in the vicinity of +/Csf1p tumor that had developed to the carcinoma stage (Fig. 3 G and H), whereas the density of F4/80+ cells was still reduced in Csf1p/Csf1p tumors, even in those that had developed to the same histological stage (Fig. 3 I and J).

These results are consistent with the hypothesis that CSF-1 acts through macrophages in its promotion of tumor progression. However, in human mammary tumor, CSF-1 receptor expression has been detected in tumor cells as well as infiltrated macrophages (8). To determine if macrophages are the only target cell at the tumor site in mice, we performed in situ hybridization for CSF-1R expression. CSF-1R–positive cells were found to be in the infiltrated cells surrounding the tumor in +/Csf1p mammary glands but the tumor cells were consistently negative (Fig. 4 A, ii). The positively stained cells were mononuclear, dendritic cells with elongated cell bodies, and were in the same location as F4/80+ cells consistent with their identification as macrophages (Fig. 4 A, iii). Very few positive cells were found when the same gms antisense probe was hybridized to tumor-bearing Csf1p/Csf1p mammary glands (data not shown). The sense gms probe on adjacent +/Csf1p mammary gland sections was consistently negative (Fig. 4 A, i).

To obtain a quantitative comparison of macrophage infiltration in the mammary tumors of Csf1p/Csf1p and +/Csf1p PyMT mice, the expression of gms was determined by Northern analysis. Expression of gms mRNA was detected in +/Csf1p mammary glands and an approximately threefold increase of the mRNA was observed at 6 wk compared with 4 wk of age (Fig. 4 B). The gms mRNA in Csf1p/Csf1p mammary glands was barely detectable until 12 wk, at which age the level of gms mRNA was less than the level found in +/Csf1p mammary glands at 4 wk of age (Fig. 4 B). At all mononuclear phagocytes express the CSF-1R, the data is consistent with both the F4/80 and gms.
2. Delineating Subfigures
Heuristic Baseline Method

1. Detect letters that denote each subfigure (‘A’, ‘B’, etc.) using connected component analysis.

2. Use a greedy tiling mechanism that places the letter in the top left corner of panels to construct a rectangular layout for each panel in a figure.

3. A figure is cut into multiple sub-panels by straight lines that go along the top or left side of each detected letter.
Baseline Example
‘You Only Look Once: Unified, Real-Time Object Detection

Redmon et al. (2015) CoRR abs/1506.02640

Used for Compound Figure Separation by Tsutsui & Crandall, (2017), ICDAR
In standard YOLO, the hyper-parameter: **# of Grids** is set as 13*13. It's too many for academic images, the network always favor tiny subfigures. (bias)

Different layout is sensible to Grid #. We apply a model based on applying multiple sets of grids.
Layout Aware YOLO
Preliminary Analysis of data from INTACT 2017

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heuristic Greedy Cut</td>
<td>0.78</td>
</tr>
<tr>
<td>YOLO</td>
<td>0.76</td>
</tr>
<tr>
<td>Layout Aware YOLO</td>
<td>0.84</td>
</tr>
</tbody>
</table>
3a. Classifying Subfigures

- LeNet image classification algorithm
- Manually annotated images from INTACT database

Table 1. Subfigure type detection performance.

<table>
<thead>
<tr>
<th>Figure Type</th>
<th>N(train)</th>
<th>N(test)</th>
<th>Tagging Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chart</td>
<td>980</td>
<td>315</td>
<td>0.92</td>
</tr>
<tr>
<td>Diagram</td>
<td>819</td>
<td>197</td>
<td>0.40</td>
</tr>
<tr>
<td>Gel</td>
<td>1402</td>
<td>404</td>
<td><strong>0.83</strong></td>
</tr>
<tr>
<td>Histology</td>
<td>1299</td>
<td>398</td>
<td>0.97</td>
</tr>
</tbody>
</table>
3b. Classifying Evidence Text

- Train simple CNN + LSTM classifiers predict detection method codes from INTACT ‘evidence fragments’ and sub-captions
- 3,366 open access INTACT records

Table 2. Accuracy for experimental type classification from text.

<table>
<thead>
<tr>
<th>Detection Method</th>
<th>Evidence Fragment</th>
<th>Sub-Caption</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LSTM</td>
<td>CNN</td>
</tr>
<tr>
<td>Participant (48 types)</td>
<td>0.37</td>
<td>0.48</td>
</tr>
<tr>
<td>Participant (6 types)</td>
<td>0.58</td>
<td>0.70</td>
</tr>
<tr>
<td>Interaction (122 types)</td>
<td>0.26</td>
<td>0.50</td>
</tr>
<tr>
<td>Interaction (18 types)</td>
<td>0.71</td>
<td>0.73</td>
</tr>
<tr>
<td>Interaction(Co-IP tagging)</td>
<td>0.79</td>
<td>0.84</td>
</tr>
<tr>
<td>Participant(WB tagging)</td>
<td>0.71</td>
<td>0.79</td>
</tr>
</tbody>
</table>
Towards Evidence Extraction Templates
Linked Data as a Research Object
http://purl.org/ske/ro/semsci18

Tools:
- https://github.com/SciKnowEngine/evidX/releases/tag/v0.1.0
  TensorFlow Classifiers used to classify text of subfigure captions by method type.
  Text preprocessing pipelines
- https://github.com/SciKnowEngine/lapdftext
  PDF image and text extraction tools

Data
- https://doi.org/10.5281/zenodo.1315036
  ‘Molecular Biology Open Access Pubmed Word and Sentence Representations’
- https://doi.org/10.5281/zenodo.13150211
  ‘Method Classification of Open Access INTACT Molecular Interaction data.’
- https://doi.org/10.5281/zenodo.1319198
  ‘Partitioned Image Data for Machine Learning Analysis of Molecular Biology Figures’
Acknowledgements

Nanjun Peng
Xiangci Li
Pradeep Dasigi
Ed Hovy
Anita De Waard

NIH 5R01LM012592-02

DARPA Big Mechanism
under ARO contract
W911NF-14-1-0436
CAN A NINJA CATCH AN ARROW? ON THIS EPISODE, WE'LL FIND OUT!

MMM, SCIENCE. HEY, MYTHBUSTERS IS ENTERTAINING, BUT IT'S NOT SCIENCE.

BRAAAAAA!! NNS...
ZOMBIE FEYNMAN! YOU GOT A PROBLEM WITH MYTHBUSTERS?
THEY FAIL AT BASIC RIGOR!

"IDEAS ARE TESTED BY EXPERIMENT." THAT IS THE CORE OF SCIENCE. EVERYTHING ELSE IS BOOKKEEPING.

BY TEACHING PEOPLE TO HOLD THEIR BELIEFS UP TO EXPERIMENT, MYTHBUSTERS IS DOING MORE TO DRAG HUMANITY OUT OF THE UNSCIENTIFIC DARKNESS THAN A THOUSAND LESSONS IN RIGOR.

SHOW THEM SOME LOVE.

ANYWAY, BACK TO ZOMBIE STUFF. I HUNGER FOR BRAAAAAA!! NNS!

UH, TRY THE PHYSICS LAB NEXT DOOR. I SAID BRAINS. ALL THEY'VE GOT ARE STRING THEORISTS.