The Subcellular Location Image Finder (SLIF)

http://slif.cbi.cmu.edu
Ultimate Goal of the field

- Machine understanding of biological journal articles (text and image)
- Criteria for success: Turing test - have machine be able to answer questions about an article as well as a human scientist
Intermediate Goal

- Extract information from combination of text and any kind of image in biological journal article
- Criteria for success: Achieve high precision and recall for extracted assertions (compared to expert scientist)
Immediate Goal (SLIF)

- Extract information about subcellular location from captions and figures containing fluorescence microscope images in biological journal articles

- Criteria for success: Achieve high precision and recall for extracted assertions (compared to expert scientist)
State of art: Bio Journal Information Extraction

- A number of systems to index literature via extracted terms
- A few systems to index image content in literature
- A few systems for document classification
Overview: Image processing tasks

- Segment into "panels"
- Detect & remove annotations
- Classify panels
- Find scale bars
Figure 1. (A) Single confocal optical section of BY-2 cells expressing U2B 0-GFP, double labeled with GFP (left panel) and autoantibody against p80 coilin (right panel). Three nuclei are shown, and the bright GFP spots colocalize with bright foci of anti-coilin labeling. There is some labeling of the cytoplasm by anti-p80 coilin. (B) Single confocal optical section of BY-2 cells expressing U2B 0-GFP, double labeled with GFP (left panel) and 4G3 antibody (right panel). Three nuclei are shown. Most coiled bodies are in the nucleoplasm, but occasionally are seen in the nucleolus (arrows). All coiled bodies that contain U2B 0 also express the U2B 0-GFP fusion. Bars, 5 m m. Movement of Coiled Bodies Vol. 10, July 1999 2299
SLIF (Subcellular Location Image Finder) automatically extracts information about protein subcellular locations from figure-caption pairs in biological literature. SLIF separates figures into panels and decides which panels contain fluorescence microscope images (FMI). It applies image processing methods to analyze the FMI and extract a quantitative description of the localization patterns they display. The associated captions are also processed to identify which portions of the caption refer to which panels and to identify the names of proteins contained in the captions. The results of this analysis are stored in the SLIF database.

Our long-term goal is to develop a large library of annotated and analyzed fluorescence microscope images, in order to support data-mining.

**PNAS, version 3.0**

The current version of the database contains records for 15180 papers from volumes 94-99 of the Proceedings of the National Academy of Sciences (USA), generously made available by the Academy for demonstration purposes.

**Pubmed Central, version 1.1**

This database includes all open access articles in Pubmed Central (approximately 42,000 articles through 30 April 2007).

A service of the Robert F. Murphy laboratory

Departments of Biological Sciences, Biomedical Engineering, and Machine Learning

and Center for Bioimage Informatics

Carnegie Mellon University, Pittsburgh, Pennsylvania, U.S.A.
SLIF (Subcellular Location Image Finder) automatically extracts information about protein subcellular locations from figure-caption pairs in biological literature. SLIF separates figures into panels and decides which panels contain fluorescence microscope images (FMI). It applies image processing methods to analyze the FMI and extract a quantitative description of the localization patterns they display. The associated captions are also processed to identify which portions of the caption refer to which panels and to identify the names of proteins contained in the captions. The results of this analysis are stored in the SLIF database, which can be queried either interactively (using the links at the left) or via external program-generated links.

Our long-term goal is to develop a large library of annotated and analyzed fluorescence microscope images, in order to support data-mining and integration with other databases.

Acknowledgments People Reference
<table>
<thead>
<tr>
<th>Caption</th>
<th>Figure</th>
<th>Panel</th>
<th>Protein</th>
<th>Cell Type</th>
<th>Location Inferred from Image</th>
<th>Location Inferred from GO Terms</th>
<th>Resolution (μm/pixel)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prevention of apoptosis in MCF-7 cells by Bcl-2 and Bcl-XL and by mutants with restricted subcellular localization. A) Cell lysates containing 20 μg of protein from vector control (neo) or transfected MCF-7 cells expressing</td>
<td><img src="image1.png" alt="Click to view in a new window" /></td>
<td>N/A</td>
<td>N/A</td>
<td>Endosome</td>
<td>N/A</td>
<td>0.057576</td>
<td></td>
</tr>
<tr>
<td>Morphology and Pax6 expression of cultured corneal epithelial cells. (A) Hematoxylin and</td>
<td><img src="image2.png" alt="Click to view in a new window" /></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Caption</td>
<td>Figure</td>
<td>Panel</td>
<td>Protein</td>
<td>Cell Type</td>
<td>Location Inferred from Image</td>
<td>Location Inferred from GO Terms</td>
<td>Resolution (μm/pixel)</td>
</tr>
<tr>
<td>---------</td>
<td>--------</td>
<td>-------</td>
<td>---------</td>
<td>-----------</td>
<td>----------------------------</td>
<td>-------------------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>TNF does not interfere with neural stem cell differentiation. (A) Immunostaining of neuronally differentiated neurosphere-derived NSCs. Cells were cultured on poly-D-lysine/laminin-coated dishes in</td>
<td>![Image of Figure A]</td>
<td>![Image of Panel]</td>
<td>![Image of Protein]</td>
<td>![Image of Cell Type]</td>
<td>![Image of Location]</td>
<td>![Image of GO Terms]</td>
<td>0.09</td>
</tr>
<tr>
<td>bFGF</td>
<td>N/A</td>
<td>Endosome</td>
<td>N/A</td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: If your browser manages only one cookie per all the windows connecting to the same server.
SLIF
Subcellular Location Image Finder

Home
Search by words in caption
Search by pixel resolution
Search by protein name
Search for FMI figures/panels
Search by location inferred from GO terms
Search by locations inferred from Images
Browse by combination of locations inferred from GO Terms and Images
Advanced Search
About SLIF
Help

Search by pixel resolution, within Captions with the word(s) tubulin.

- Search by pixel resolution

The min resolution is 12 Microns/Pixel and the max is 0.05 Microns/Pixel

- High: 0.05 ~ 0.311
- Medium: 0.311 ~ 1.931
- Low: 1.931 ~ 12
- Other: ____________ ~ ____________

Number of records displayed per page: 10

Search
Subcellular Location Image Finder

Search by protein name, within Images with pixel resolution between 0.05 and 0.31072325059538597 microns/pixel, within Captions with the word(s) tubulin.

- Protein names extracted by CRFs
- Protein names extracted by Sprot matcher
- Protein names extracted by Uniref100 matcher

Search by protein name
- Protein names extracted by CRFs
- Protein names extracted by Sprot matcher
- Protein names extracted by Uniref100 matcher

Number of records displayed per page: 10

Note: If your browser manages only one cookie per all the windows connecting to the same server.
Search by protein name, within
Images with pixel resolution between 0.05 and 0.31072325059538597 microns/pixel, within
Captions with the word(s) tubulin.

- Protein names extracted by CRFs

---Please select---
BFGF
Dissociated NSCs
EGF
GFAP
II
TNF

- Protein names extracted by Uniref100 matcher

---Please select---

Search by protein name
- Protein names extracted by CRFs

---Please select---

- Protein names extracted by Sprot matcher

---Please select---

- Protein names extracted by Uniref100 matcher

---Please select---

Number of records displayed per page: 10
Subcellular Location Image Finder

Search by protein name, within Images with pixel resolution between 0.05 and 0.31072325059538597 microns/pixel, within Captions with the word(s) tubulin.

- Protein names extracted by CRFs
- Protein names extracted by Sprot matcher
- Protein names extracted by Uniref100 matcher

Search by protein name

- Protein names extracted by CRFs
- Protein names extracted by Sprot matcher
- Protein names extracted by Uniref100 matcher

Number of records displayed per page: 10

Note: If your browser manages only one cookie per all the windows connecting to the same server.
Search by protein name, \textit{within}
Images with \textbf{pixel resolution} between 0.05 and 0.31072325059538597 microns/pixel, \textit{within}
Captions with the word(s) \textit{tubulin}.

- Protein names extracted by CRFs
- Protein names extracted by Sprot matcher
- Protein names extracted by Uniref100 matcher

Search by protein name

- Protein names extracted by CRFs
- Protein names extracted by Sprot matcher
- Protein names extracted by Uniref100 matcher

Number of records displayed per page: 

Note: If your browser manages only one cookie per all the windows connecting to the same server.
Search for:

**Protein name(s) GFAP, within**

Images with **pixel resolution** between 0.05 and 0.31072325059538597 microns/pixel, **within** Captions with the word(s) tubulin.

Results 1 - 7 of 7

<table>
<thead>
<tr>
<th>Caption</th>
<th>Figure</th>
<th>Panel</th>
<th>Protein</th>
<th>Cell Type</th>
<th>Location Inferred from Image</th>
<th>Location Inferred from GO Terms</th>
<th>Location (μm/pixel)</th>
<th>FMI and Protein Name Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>TNF does not interfere with neural stem cell differentiation (A)</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td>GFAP</td>
<td>N/A</td>
<td>Endosome</td>
<td>N/A</td>
<td>0.1</td>
<td>0.9823</td>
</tr>
<tr>
<td>Immunostain of neurally differentiated neurosphere-derived NSCs. Cells were cultured</td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Click to view in a new window.*
Subcellular Location Image Finder

Search by location inferred from GO terms

GO terms found for protein names extracted by: Sprot

Subcellular Location

Number of records displayed: 100

--- Please select ---
Chromosome
Cytoplasm
Cytoplasmic Vesicle
Cytoskeleton
Cytosol
Endosome
ER
Extracellular
Golgi apparatus
Intracellular
Membrane
Mitochondrion
Nuclear Membrane
Nucleoplasm
Nucleus
Other Component
Ribosome
Undecidable
Search for:

**Protein name extractor(s)** Sprot, *within*

**Locations** inferred from the GO term **Golgi apparatus**.

Results 1 - 41 of 41

<table>
<thead>
<tr>
<th>Link to PIR Query</th>
<th>Link to SLIF Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANK</td>
<td>Link to SLIF query: ANK</td>
</tr>
<tr>
<td>BSA</td>
<td>Link to SLIF query: BSA</td>
</tr>
<tr>
<td>CDK1</td>
<td>Link to SLIF query: CDK1</td>
</tr>
<tr>
<td>COX</td>
<td>Link to SLIF query: COX</td>
</tr>
<tr>
<td>COX-2</td>
<td>Link to SLIF query: COX-2</td>
</tr>
<tr>
<td>CP</td>
<td>Link to SLIF query: CP</td>
</tr>
<tr>
<td>DP</td>
<td>Link to SLIF query: DP</td>
</tr>
<tr>
<td>EGF</td>
<td>Link to SLIF query: EGF</td>
</tr>
<tr>
<td>F1</td>
<td>Link to SLIF query: F1</td>
</tr>
<tr>
<td>F2</td>
<td>Link to SLIF query: F2</td>
</tr>
<tr>
<td>FGF-1</td>
<td>Link to SLIF query: FGF-1</td>
</tr>
<tr>
<td>GAPDH</td>
<td>Link to SLIF query: GAPDH</td>
</tr>
<tr>
<td>HEK</td>
<td>Link to SLIF query: HEK</td>
</tr>
<tr>
<td>HSS</td>
<td>Link to SLIF query: HSS</td>
</tr>
<tr>
<td>HSulf-2</td>
<td>Link to SLIF query: HSulf-2</td>
</tr>
<tr>
<td>HTH</td>
<td>Link to SLIF query: HTH</td>
</tr>
<tr>
<td>HuR</td>
<td>Link to SLIF query: HuR</td>
</tr>
</tbody>
</table>
Subcellular Location Image Finder

- Search by words in caption
- Search by pixel resolution
- Search by protein name
- Search for FMI figures/panels
- Search by location inferred from GO terms
- Search by locations inferred from Images
- Browse by combination of locations inferred from GO Terms and Images
- Advanced Search
- About SLIF
- Help

- Search by pixel resolution, within
  Images with **pixel resolution** between 0.2 and 1.0 microns/pixel.

- Search by pixel resolution

The min resolution is 12 Microns/Pixel and the max is 0.05 Microns/Pixel

- High: 0.05 ~ 0.311
- Medium: 0.311 ~ 1.931
- Low: 1.931 ~ 12
- Other: 0.2 ~ 1.0

Number of records displayed per page: [10]
### Expression of HDGF family members in granular cells of the cerebellum.

Antibodies against HDGF (A), HRP-2 (D), HRP-3 (G) and NeuN (B, E, H) were used to detect the proteins in brain slices of adult mice. Double stainings demonstrate that in

<table>
<thead>
<tr>
<th>Caption</th>
<th>Figure</th>
<th>Panel</th>
<th>Protein</th>
<th>Cell Type</th>
<th>Location Inferred from Image</th>
<th>Location Inferred from GO Terms</th>
<th>Resolution (μm/pixel)</th>
<th>FMI Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression of HDGF family members in granular cells of the cerebellum.</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td>HDGF</td>
<td>N/A</td>
<td>ER</td>
<td>Cytoplasm</td>
<td>0.2375</td>
<td>0.9996</td>
</tr>
<tr>
<td>Antibodies against HDGF (A), HRP-2 (D), HRP-3 (G) and NeuN (B, E, H) were used to detect the proteins in brain slices of adult mice. Double stainings demonstrate that in granular cells of the cerebellum.</td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Subcellular Location Image Finder

- Home
- Search by words in caption
- Search by pixel resolution
- Search by protein name
- Search for FMI figures/panels
- Search by location inferred from Go terms
- Search by locations inferred from Images
- Browse by combination of locations inferred from GO Terms and Images
- Advanced Search
- About SLIF
- Help

Generate Location Comparison Matrix, within Images with pixel resolution between 0.2 and 1.0 microns/pixel.

Generate Location Comparison Matrix
## Subcellular Location Image Finder

### GO Terms vs Images

<table>
<thead>
<tr>
<th>GO Term</th>
<th>Endosome</th>
<th>ER</th>
<th>Golgi</th>
<th>Lysosome</th>
<th>Mitochondria</th>
<th>Nucleolus</th>
<th>Tubulin</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytoplasm</td>
<td>78</td>
<td>23</td>
<td>1</td>
<td>1</td>
<td>20</td>
<td>2</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Cytoskeleton</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Cytosol</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ER</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Extracellular</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Golgi apparatus</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Membrane</td>
<td>18</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Mitochondrion</td>
<td>15</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Nucleus</td>
<td>81</td>
<td>27</td>
<td>1</td>
<td>1</td>
<td>27</td>
<td>0</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Other Component</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Undecidable</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Subcellular Location Image Finder

Searched for:

**Locations** inferred from the GO term **Mitochondrion**, *within*
**Location(s)** inferred from the images **Mitochondria**, *within*
Images with **pixel resolution** between 0.2 and 1.0 microns/pixel.

Results 1 - 10 of 32

<table>
<thead>
<tr>
<th>Caption</th>
<th>Figure</th>
<th>Panel</th>
<th>Protein</th>
<th>Cell Type</th>
<th>Location Inferred from Image</th>
<th>Location Inferred from GO Terms</th>
<th>Resolution (µm/pixel)</th>
<th>FMI Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Muscle Spindles in Gastrocnemius Muscle and TrkA/TrkC Staining in the Tibial Nerve at P0(α) NF-M (red) and S46 (green) immunostaining in cross section of gastrocnemius muscle at P0. There are no muscle spindles</td>
<td><img src="image1.png" alt="Figure" /></td>
<td><img src="image2.png" alt="Panel" /></td>
<td>E17</td>
<td>N/A</td>
<td>Mitochondria</td>
<td>Mitochondrion</td>
<td>0.546667</td>
<td>0.9868</td>
</tr>
</tbody>
</table>

Click to view in a new window.
## Subcellular Location Image Finder

Searched for:
- **Locations** inferred from the GO term **Mitochondrion**, within
- **Location(s)** inferred from the images **Mitochondria**, within
- Images with **pixel resolution** between 0.2 and 1.0 microns/pixel.

Results 11 - 20 of 32

<table>
<thead>
<tr>
<th>Caption</th>
<th>Figure</th>
<th>Panel</th>
<th>Protein</th>
<th>Cell Type</th>
<th>Location Inferred from Image</th>
<th>Location Inferred from GO Terms</th>
<th>Resolution (µm/pixel)</th>
<th>FMI Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Immunohistochemical localization of OCTs in murine bronchi. OCT1-immunolabelling is localized to the apical membrane of ciliated epithelial cells in wild-type FVB mice (arrows in A). The specificity of this labelling is</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td>OCT1</td>
<td>N/A</td>
<td>Mitochondria</td>
<td>Mitochondrion</td>
<td>0.75</td>
<td>0.9115</td>
</tr>
</tbody>
</table>
Advanced Search

- Search by words in caption
- Search by protein name
- Search for FMI figures/panels
- Search by location inferred from GO terms
- Search by locations inferred from Images
- Browse by combination of locations inferred from GO Terms and Images
- Advanced Search
- About SLIF
- Help

Murphy Lab SLIF Service

SLIF Web Application Version 2.1
Database: pmc

Note: If your browser manages only one cookie use all the volunteered information in the same cookie.

The min resolution is 12 Microns/Pixel and the max is 0.05 Microns/Pixel
Linking to SLIF from another website

- Queries against the database can be made from outside the website using http://slif.cbi.cmu.edu/PMC/search.jsp?arguments

- The arguments are:
  - protein=<protein name>
  - level=figure OR level=panel
  - type=FMI (NOTE that BOTH level and type must be present if either is present)
  - pixel_size_lo=<lower bound>
  - pixel_size_hi=<upper bound> (NOTE that both upper and lower bounds must be specified)
  - location=<subcellular location>
SOAP interface - query DTD

<!ELEMENT slif_query (protein_name, fmi_level, pixel_res_lower_bound, pixel_res_upper_bound, subcellular_location)>
<!ELEMENT protein_name (#PCDATA)>
<!ELEMENT fmi_level EMPTY>
<!ATTLIST fmi_level figure_or_panel (figure|panel) #REQUIRED>
<!ELEMENT pixel_res_lower_bound (#PCDATA)>
<!ELEMENT pixel_res_upper_bound (#PCDATA)>
<!ELEMENT subcellular_location (#PCDATA)>
SLIF web client options

- gui  launches GUI search utility
- p    protein name
- il   image level (figure or panel)
- go   GO location
- hr   highest image resolution
- lr   lowest image resolution
- id   article id
- w    words in caption
```xml
<?xml version="1.0"?>
<!DOCTYPE slif_query_results SYSTEM "http://slif.cbi.cmu.edu/slif_result.dtd">
<slif_query_response>
<query>
<fmi_level>panel</fmi_level>
<scale>0.2,2.0</scale>
</query>
<result_summary>
<num_records>8202</num_records>
</result_summary>
<query_results>
<slif_result>
<caption>...
```

Carnegie Mellon
java -jar SLIFPMCSearh.jar -lr 0.2 – hr 2.0 -il panel

<slif_result>
<caption>notch and delta gene expression in zebrafish …</caption>
<figure_url>p005-paper/paper4439-figure/figure0001-image_box.jpg</figure_url>
<panel_url>p005-paper/paper4439-figure/figure0001-panel/panel_6_green-image.jpg</panel_url>
<protein_name>n1b</protein_name>
<cell_name>N/A</cell_name>
<subcellular_location>Endosome</subcellular_location>
<pixel_resolution>1.6</pixel_resolution>
</slif_result>
java -jar SLIFPmCSearch.jar –p cadherin –il panel

<?xml version="1.0"?>
<!DOCTYPE slif_query_results SYSTEM "http://slif.cbi.cmu.edu/slif_result.dtd">
<slif_query_response>
<query>
<protein>cadherin,</protein>
<fmi_level>panel</fmi_level>
</query>
<result_summary>
<num_records>8</num_records>
</result_summary>
<query_results>
<slif_result>
Summary

- Complex pipeline for parsing figure and caption pairs
- Web pages to provide interactive access to results of analysis
- Web service for programmatic access
- Work continues…
Acknowledgments

Commonwealth of Pennsylvania Tobacco Settlement Funds
NIH R01 GM078622

Past:
Meel Velliste, Greg Porreca, Zhenzhen Kou, Juchang Hua

Current:
Amr Ahmed, Andrew Arnold, Luis Coelho, Saboor Sheikh

Carnegie Mellon