Image content-based searching and tools for image-derived modeling

Robert F Murphy
Ray & Stephanie Lane Professor of Computational Biology and Professor of Biological Sciences, Biomedical Engineering and Machine Learning
External Senior Fellow, Albert Ludwig University of Freiburg
Content-based image search

- Also called Query-by-image-content (QBIC)
- Find images whose content, as reflected by image features, is similar to one or more query images
- Can use positive and/or negative examples
- Can be iterative (relevance feedback)
Why content-based search?

- In compound screening, to see whether anyone else in the organization (or the world) has seen a particular pattern before
- In proteomics, to find other proteins that show a subcellular location that is difficult to assign
OMERO.searcher: content-based image search for microscope images

Baek Hwan Cho, Ivan Cao-Berg, Jennifer Ann Bakal & Robert F Murphy

Affiliations | Corresponding author

Published online 28 June 2012

To the Editor:

Fluorescence microscopy is growing dramatically both in terms of technical capabilities and the volume of images generated. Online repositories have been created to provide public access to images and opportunities for joint research for many scientists¹. This has reintroduced challenges faced when sequence and structure databases were being established: developing fast and
Testing search effectiveness

- Combined sets of images of different patterns into database
- Searched with one or more images of one of the patterns
- Observed how highly ranked images of that pattern are relative to other patterns (measure area under ROC curve)
Results for images from diverse image sources

Images from The Cell library
Results for different patterns from same source
OMERO.searcher

• Open source content-based image database search
• Three ways to use:
  1. Server – install on top of your OMERO db
  2. External Search – search via web page
  3. Local Client – search via Python script

Can search with local images on any computer
1. OMERO.searcher Server

- Install on top of OMERO 4.3
- Installs feature calculation, content search and changes to OMERO.web
OMERO.searcher workflow

a

1. Image Importing via Importer
2. Select Image(s)
3. Content Search
4. Select Additional Image(s)
5. Satisfied?
   - Yes: End
   - No: Feature Calculation, Content DB Update, automatically executes

b

- # retrieved images 10
- Search by image content
- Search results using image content
- Query configuration
- Selected features and dataset
- Go to Basket

- Neg
- Pos
2. External Search page

- Django app installed with Server
- Allows search using external images

OMERO.searcher 1.1.2

Search OMERO database for similar images using local files

This page permits content-based searching of an OMERO database using OMERO.searcher using local image files. The images must be in a format readable by the Python Imaging Library (which includes most standard image formats like TIFF, PNG, and JPEG), and must contain a single image channel. The two images must also be of the same size and resolution. Images are read by the server, features are calculated and compared to the database, and a list of the most similar images is returned. Images and features are not stored on the server.
2. External Search results

**OMERO.searcher 1.1.2 Results**

<table>
<thead>
<tr>
<th>Query Image(s)</th>
<th>110509RBTRBT_GQ2__H___6_T_001_ch_01_image_00015_Z_001.bmp</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>110509RBTRBT_GQ2__H___6_T_001_ch_00_image_00015_Z_001.bmp</td>
</tr>
<tr>
<td>Content Database</td>
<td>/usr0/local/omero.server/OMERO.server-Beta-4.3.3/lib/python/omero/web/searcher2/contentDB/103_all_sls34_content_db_0000000000000000000001.pkl</td>
</tr>
<tr>
<td>Original Scale</td>
<td>0.16125</td>
</tr>
<tr>
<td>Comparison Scale</td>
<td>0.16125</td>
</tr>
<tr>
<td>Feature Set</td>
<td>sls34</td>
</tr>
<tr>
<td>Number to Retrieve</td>
<td>10</td>
</tr>
</tbody>
</table>

*Results* (if images are not displayed, click on an info link to log into the OMERO database - the username is demo_nm and the password nature@u0816 - use the back button on your browser to return to this page and then hit refresh)
3. Local Client

- Runs on local computer
- Requires python and packages
- Returns HTML page

```
python omero.searcher.py *c1*.bmp *c0*.bmp 0.2 10
```
Local Client version 1.2 released January 8, 2013

- Supports user-specified feature sets
- Supports non-OMERO databases
User-defined feature sets

• Define a feature set as an ordered set of individual features
• Use python template to produce a function that calculates desired features
Portable content database

- Contains feature values (for one feature set) for each image in database, with URLs linking to that image
- Native format is python .pkl file
- Support for converting to/from .hd5 files
- Allows embedding of user-defined feature calculation code
Making a database available for search

• If OMERO database, install OMERO.searcher
• If non-OMERO database, create content DB file containing features and URLs
• In either case, register the content DB so that other users can access, or let people know URL for database in some other manner
Local Client workflow

- User specifies database URL, feature set name, query images
- Program
  - looks for content DB for specified DB (and checks if it is latest version)
  - runs feature calculation code on the query images (if features not already calculated)
  - produces HTML file showing results of search
Publicly searchable databases

• Released
  – Nature Methods demo1
  – **PSLID RandTag Release 2 (new)**
    • [http://pslid.org](http://pslid.org)
    • 6099 CD-tagged clones, 99955 HTM images, 2777 confocal images, 2937 genomic sequences
    • Identified 116 tagged proteins in 189 clones with high confidence
    • Identified additional 70 tagged proteins in 133 clones with medium confidence
PSLID - Protein Subcellular Location Image Database

PSLID is both a web service providing access to collections of images showing subcellular location patterns and an open-source image database software system. There are two main collections: the original Public database containing various small image sets useful for training and testing various automated analysis approaches, and the RandTag database containing images of NIH 3T3 clones expressing randomly-tagged proteins.

Go to PSLID RandTag Release 2 (released January 9, 2013)

Go to the PSLID Public Database Release 4 (released August 1, 2008)

Download datasets contained within PSLID

Download SLIC feature calculation and pattern analysis software used in PSLID and full PSLID software

Go to PSLID RandTag Release 1 (released July 1, 2009)

Supported by the National Institute of General Medical Sciences, Grant GM075205
**actin related protein 2/3 complex, subunit 2**

- **Gene:** ARPC2_MOUSE
- **Location from external databases:**
  - Uniprot: cell leading edge; focal adhesion; cell projection; arp2/3 protein complex
  - MGI: golgi apparatus; cytoplasm; cell leading edge; focal adhesion; arp2/3 protein complex; cytoskeleton; cell projection
- **Location from image classification:**
  - Major (minor): cytoplasm (membrane, no apparent fluorescence); (cytoplasm, membrane, no apparent fluorescence)
- **Current Curated Location:** none listed
- **Proposed Location:**
  - Proposed Location:
  - Current Proposed Location(s): none listed
  - Proposed Location:
  - Submit proposed location

### EM1D8_IC100
- **Location from image classification:**
  - Major class(es): cytoplasm
  - Minor class(es): membrane, no apparent fluorescence
- **Cell name:** 3T3-Ctag-EM1D8
- **Experiment title:** 3T3-Ctag-EM1
- **Sequence information:**
- **Gene assignment confidence:** Low

### FX1B4_IC100
- **Location from image classification:**
  - Major class(es): cytoplasm
  - Minor class(es): membrane, no apparent fluorescence
- **Cell name:** 3T3-Ctag-FX1B4
- **Experiment title:** 3T3-Ctag-FX1
- **Sequence information:**
- **Gene assignment confidence:** Low

### FX1E9_IC100
- **Location from image classification:**
  - Major class(es): cytoplasm
  - Minor class(es): membrane, no apparent fluorescence
- **Cell name:** 3T3-Ctag-FX1E9
- **Experiment title:** 3T3-Ctag-FX1
- **Sequence information:**
- **Gene assignment confidence:** High

### FX2B3_IC100
- **Location from image classification:**
  - Major class(es): Minor class(es): cytoplasm, membrane, no apparent fluorescence
- **Cell name:** 3T3-Ctag-FX2B3
- **Experiment title:** 3T3-Ctag-FX2
- **Sequence information:**
- **Gene assignment confidence:** High
URL-based searching of PSLID

• Search by target or clone
• Return HTML or XML
• Can download images matching a query as a ZIP file of OME-TIFF images
Planned publicly searchable databases

- The Cell library FM Subset1 (should be released next week)
- Human Protein Atlas confocal images
OMERO.searcher plans

• Support for
  – OMERO 4.4 (done?)
  – Search across multiple servers
  – BioFormats for reading images
  – Other feature sets
  – Combining content and context for search
  – Integration into base OMERO system

• Additional OMERO databases available for public search
PatternUnmixer

• Performs supervised unmixing of subcellular location

• Versions
  – Matlab source

• Supports training on or unmixing images from local disk or multiple OMERO servers
• **Lysotracker**
• Mitotracker
• Mixture of Lysotracker and Mitotracker
PatternUnmixer plans

- Django app running on server
  - No requirements on local computer
- Heatmap displays and hit finders
- Other local feature detectors
- Add unsupervised unmixing (have code already)
CellOrganizer

- Open-source system for learning generative models of cell organization
- Version 1.8 released tomorrow
  - Includes diffeomorphic models for generation of cell and nuclear shapes
April 13, 2012: Integrated 2D/3D version 1.6 released!

The CellOrganizer project provides tools for

- learning generative models of cell organization directly from images
- storing and retrieving those models in XML files
- synthesizing cell images (or other representations) from one or more models

Model learning captures variation among cells in a collection of images. Images used for model learning and instances synthesized from models can be two- or three-dimensional static images or movies.

Version 1.6 of CellOrganizer can learn models of

- cell shape
- nuclear shape
- chromatin texture
- vesicular organelle size, shape and position
- microtubule distribution.

These models can be conditional upon each other. For example, for a given synthesized cell instance, organelle position is dependent upon the cell and nuclear shape of that instance.

Cell types for which generative models for at least some organelles have been built include human HeLa cells, mouse NIH 3T3 cells, and Arabidopsis protoplasts. Planned projects include mouse T lymphocytes and rat PC12 cells.

Support for CellOrganizer has been provided by grants GM075205 and GM090033 from the National Institute of General Medical Sciences, by a Forschungspreis from the Alexander von Humboldt Foundation, and by the School of Life Sciences of the Freiburg Institute for Advanced Studies.
Statistical modeling and generation

Many cell images → extract parameters → statistical model → randomly sample new parameters → synthetic cell images
3D Nuclear Shape – Cylindrical Spline Surface

33 parameters (32 spline coefficients + height)
Diffeomorphic analysis of shape

- Sometimes cell or nuclear shapes are irregular
- Can use distance between shapes to characterize shape instead of parameters of model (based on work by Michael Miller and colleagues)

Gustavo Rohde
Morphing one shape into another

Starting shape

Target shape

Distance

0  0.0165  0.0191  0.0194  0.0195
Shape space

- Can measure distance between all pairs of shapes to construct a “shape space”
- Captures essential aspects of how shapes vary
- Can be applied to nuclei, cells, organelles, etc.
- Can be applied in 2D, 3D, 4D and to more than one component at a time
Cell and nuclear shape space
CellOrganizer vs. traditional HCS

Traditional HCS

Differences

Features
Control images from HCS1
Params
Differences
Features
Hit images from HCS1
Params
Differences
Features
Control images from HCS2
Params
Differences
Features
Hit images from HCS2
Params

not comparable
comparable
CellOrganizer
A Modest Proposal

- Register the large collection of cellular images within this consortium through generative models