Published in final edited form as:
Nat Methods. ; 9(7): 633–634. doi:10.1038/nmeth.2086.

OMERO.searcher: Content-based image search for microscope images

Baek Hwan Cho[1], **Ivan Cao-Berg**[1], **Jennifer Ann Bakal**[2], and **Robert F. Murphy**[1,2,3,4,5]

[1]Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

[2]Center for Bioimage Informatics, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

[3]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

[4]Department of Machine Learning, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

[5]Freiburg Institute for Advanced Studies, Albert Ludwig University of Freiburg, Germany

Fluorescence microscopy is growing dramatically both in terms of technical capabilities and the volume of images generated. A number of online repositories have been created to begin providing public access to images and opportunities for joint research to many scientists.[1] This has reintroduced challenges faced when sequence and structure databases were being established: developing fast and effective means of searching for records (images) either by context (e.g., what protein is labeled) or content (e.g., what pattern it displays). Image databases normally contain context descriptors in the form of annotations describing the source of the sample, the protocol used to prepare it, the instrumental settings used, and the laboratory that produced it. Searches can readily be done on one or more of these annotations, although incomplete or inconsistent annotation remains a problem. Searching for images based on their contents is much less developed. Some content annotations may be provided in the form of labels (such as Gene Ontology terms) resulting from either visual or automated analysis, and therefore images can be retrieved using them in the same way as context terms. However, these are limited by the “resolution” of the terms used and do not facilitate discovery of new patterns or similarities between known patterns that were not previously recognized. Content-based image retrieval (also known as Query-by-image content) was proposed many years ago to address this issue; this method takes one or more images as a query and retrieves the most similar images in terms of numerically computed features.[2] However, current fluorescence microscopy image databases do not provide these methods. Here we present a content-based image searcher for microscope images, OMERO.searcher (http://murphylab.web.cmu.edu/software/searcher), that can be used with any OMERO database (http://openmicroscopy.org).[3]

Corresponding author: Robert F. Murphy (murphy@cmu.edu).

Author Contributions: B.H.C. and J.A.B. performed research and contributed code, R.F.M. conceived and guided research, I.C. contributed code, and B.H.C and R.F.M wrote the manuscript.

Competing Financial Interests: The authors declare no competing financial interests.

The two requirements for content-based retrieval are a set of numerical features to describe each image and a method for combining them to measure similarity. OMERO.searcher by default uses the subcellular location features (SLFs)[4] which have previously been successfully used to identify protein location patterns in fluorescence microscope images, but these can be replaced with any numerical feature set the user devises for their own purposes (one of the advantages of the SLFs is that they are designed to be applicable to images taken at different resolutions or with different modalities). Images are ranked by their similarity to one or more query images using a modified implementation of the FALCON algorithm[5] that has been used in the Protein Subcellular Location Image Database (http://pslid.org).[6] The searcher is implemented on top of the OMERO web client service with minimum modification of its default web pages. The features for individual images are stored as an attached HDF5 file; the code can be configured to automatically calculate and store these features when a new image is uploaded to the server (or they can be calculated on demand through the web interface). The features for the entire database are also stored in one master file to facilitate fast searches. For each query, the searcher retrieves the features for the query images as well as the features for the entire database and performs a similarity measurement. Both positive and negative examples can be included in a query.

A typical work flow using OMERO.searcher is shown in Supplementary Fig. 1. After uploading images, features are calculated and stored in the database. These features are calculated at different image resolutions. A search can then be done simply by selecting one or more images and clicking the magnifying glass icon. The system automatically chooses, based on the resolution of the query images, the set of features to use. The query information is displayed on the left side of the resulting web page, and the most similar images retrieved are shown on the right. A user can refine the search by choosing images from the results, marking them as positive (retrieve more images similar to these) and negative (exclude images similar to these), and repeating until satisfied. A standalone client that does not require a local copy of OMERO is also available. It permits users to choose images on their local computer, calculate features, and find similar images in remote databases that have OMERO.searcher installed (Supplementary Note). (The next release of OMERO.searcher will support searching across multiple OMERO databases at different locations, assuming access rights.)

To test how well the searcher retrieves relevant images, we performed tests using two distinct fluorescence microscopy databases. Classes of images with the same content annotations were created, and images were ranked by similarity to one or more query images drawn from one of those classes (Supplementary Methods). Success was measured using the area under a receiver operating characteristic curve, in which retrieval rates for images from the desired class are compared to those for images from undesired classes. Good results were obtained for many different patterns from both databases (Fig. 1). Note that this was true even though the Cell Library contained images of different resolutions and microscope types. The quality of the results was improved by increasing the number of images in the query, and using both positive and negative examples improved the results for the same total number of labeled images (Fig. 1b). The images used in this second test were collected at 40× magnification. Very similar results were obtained when searching with downsampled versions to simulate a query with images collected at 10× magnification (Supplementary

Fig. 2). Feature sets are also available to permit searching with three-dimensional images and time series.

In conclusion, OMERO.searcher is an open-source content-based image search tool for the cell and computational biology community. It has a number of useful applications, such as asking if someone has previously observed a pattern similar to an unrecognized one, or finding examples of a particular pattern in other cell types or different modes of microscopy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by NIH grants GM075205, EB008516, and GM092708 and by grant 095931 from the Wellcome Trust. B.H.C was supported by a postdoctoral fellowship from the Korea Research Foundation Grant (KRF-2008-D00316). We thank K. Eliceiri, J. Swedlow, J. Moore, D. Orloff, L. Wu and C. Faloutsos for helpful discussions.

References

1. Swedlow JR. Nat Cell Biol. 2011; 13:183. [PubMed: 21364564]
2. Faloutsos C, et al. J Intell Inf Syst. 1994; 3:231–262.
3. Allan C, et al. Nat Meth. 2012; 9:245–253.
4. Glory E, Murphy RF. Dev Cell. 2007; 12:7–16. [PubMed: 17199037]
5. Wu L, Faloutsos C, Sycara KP, Payne TR. Proc VLDB. 2000:297–306.
6. Huang K, Lin J, Gajnak JA, Murphy RF. Proc IEEE Intl Symp Biomed Imaging. 2002:325–328.

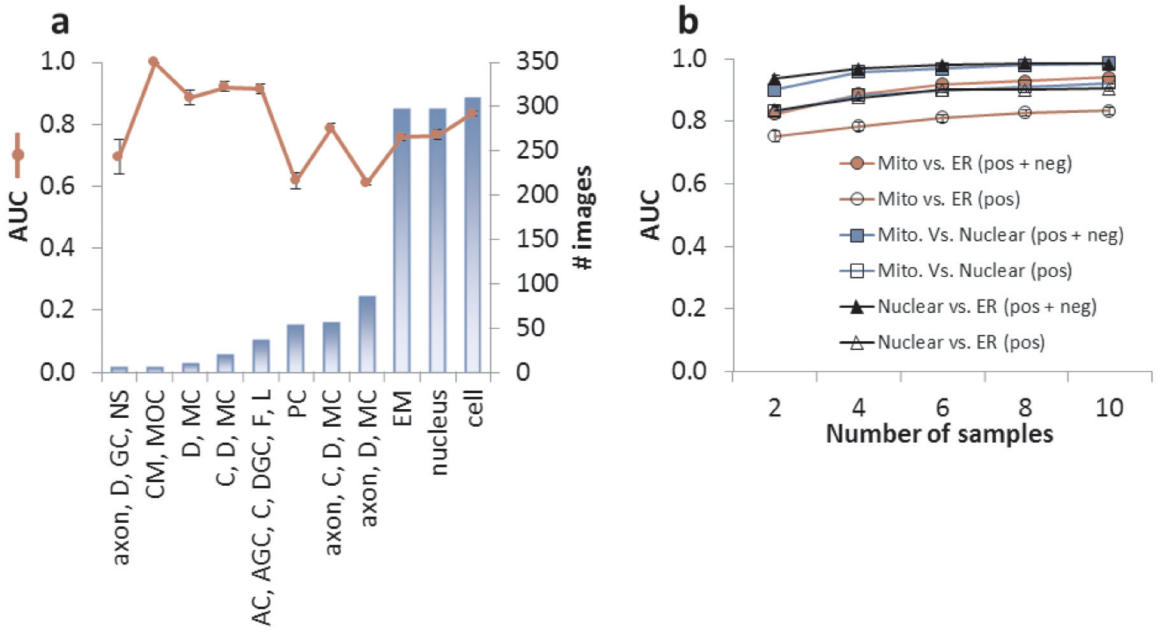

Fig. 1.

Results of retrieval performance tests. (a) Images from The Cell Library were grouped by their annotations and used to search for similar images. The area under receiver operating characteristic curves (AUC) was calculated, where a value of 1 means that every image in the same group is ranked above all images in other groups and a value of 0.5 corresponds to random ranking. The annotations are AC: actin cytoskeleton; AGC: axonal growth cone; C: cytoskeleton; CM: cytoplasmic microtubule; D: dendrite; DGC: dendritic growth cone; EM: extracellular matrix part; F: filopodium; GC: growth cone; L: lamellipodium; MC: microtubule cytoskeleton; MOC: microtubule organizing center; NS: neuron spine; PC: primary cilium. The average AUC across all patterns was 0.77. (b) A similar test was done with RandTag images from the PSLID repository, each of which was annotated with one of three protein location pattern class labels. AUC values were calculated for searches with only positive images (open symbols) or an equal mix of positive and negative images (filled symbols). The average AUC for 10 images (5 positive and 5 negative) was 0.976.