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

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 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 (such as which protein is labeled) or content (such as which pattern it displays). Image databases normally contain context descriptors in the form of annotations that describe the source of the sample, the protocol used to prepare it, the instrument settings used and the laboratory where it was produced. Searches can readily be done on one or more of these annotations, but 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 of 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 search 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].

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[4] that have been used previously to identify protein location patterns in fluorescence microscopy images, but these can be replaced with any numerical feature set the user devises for his own purposes (one advantage of the subcellular location features is that they are 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 (PSLID; 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 Figure 1**. After images are uploaded, 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

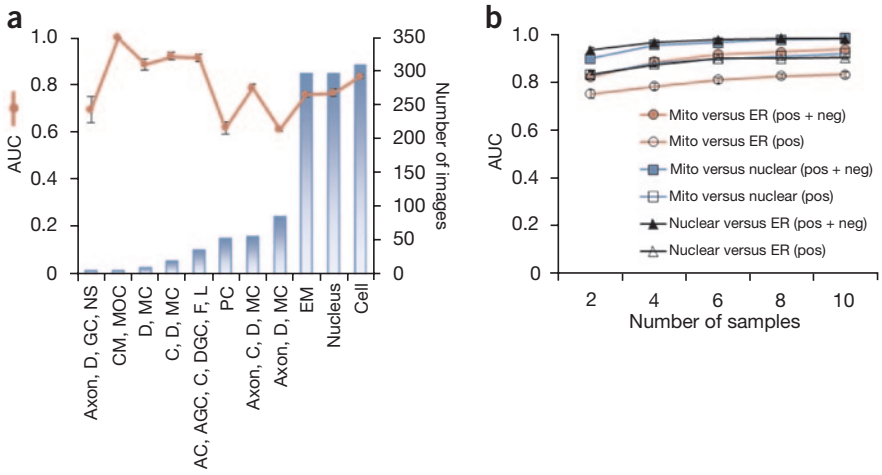

Figure 1 | Results of retrieval performance tests. (**a**) Images from The Cell were grouped by their annotations and used to search for similar images. The area under receiver operating characteristic curves (AUC) was calculated, where a value of 1 means that every image in the same group is ranked above all images in other groups, and a value of 0.5 corresponds to random ranking. 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; and 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 positive images only (pos) or an equal mix of positive and negative images (pos + neg). The average AUC for 10 images (5 positive and 5 negative) was 0.976. Mito, mitochondria; ER, endoplasmic reticulum.

search by choosing images from the results, marking them as positive (meaning 'retrieve more images similar to these') and negative ('exclude images similar to these') and repeating until satisfied. A stand-alone client that does not require a local copy of OMERO is also available (**Supplementary Software**). 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, PSLID and The Cell: An Image Library (The Cell; http://www.cellimagelibrary.org/). We created classes of images with the same content annotations and ranked the images by similarity to one or more query images drawn from one of those classes (**Supplementary Methods**). We measured success 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. We obtained good results for many different patterns from both databases (**Fig. 1**) even though The Cell contained images captured at different resolutions and from different microscope types. Increasing the number of images in the query improved result quality, as did using both positive and negative examples for the same total number of labeled images (**Fig. 1b**). The images used in this second test were collected at 40× magnification. We obtained similar results 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.

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

Note: Supplementary information is available at http://www.nature.com/doifinder/10.1038/nmeth.2086.

ACKNOWLEDGMENTS

This research was supported in part by US National Institutes of Health 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.

AUTHOR CONTRIBUTIONS

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

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Baek Hwan Cho[1], Ivan Cao-Berg[1], Jennifer Ann Bakal[2] & Robert F Murphy[1–5]

[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. e-mail: murphy@cmu.edu

1. Swedlow, J.R. *Nat. Cell Biol.* **13**, 183 (2011).
2. Faloutsos, C. *et al. J. Intell. Inf. Syst.* **3**, 231–262 (1994).
3. Allan, C. *et al. Nat. Methods* **9**, 245–253 (2012).
4. Glory, E. & Murphy, R.F. *Dev. Cell* **12**, 7–16 (2007).
5. Wu, L., Faloutsos, C., Sycara, K.P. & Payne, T.R. in *Proc. 26th Int. Conf. Very Large Data Bases* (eds., Abbadi, A.E. *et al.*) 297–306 (Morgan Kaufmann, 2000).
6. Huang, K., Lin, J., Gajnak, J.A. & Murphy, R.F. in *Proc. 2002 IEEE Int. Symp. Biomed. Imaging*, 325–328 (2002).

SimuCell: a flexible framework for creating synthetic microscopy images

To the Editor: Advances in high-content fluorescence microscopy have driven the development of analytical approaches for extracting meaningful information from rich and complex biological image data. Algorithm development can be aided dramatically by using curated test data. To evaluate the generality and performance of new algorithms, test data should contain annotation on how images differ in terms of cell phenotypes, population heterogeneity and/or microenvironmental[1] effects. Currently there is a paucity of diverse, well-annotated data. A complementary approach is to use synthetically generated data, in which biological[1] and imaging[2] effects can be varied independently and 'ground truths' are known. Although approaches exist for rendering realistic cells[3,4], creating biologically realistic cell-population images has remained challenging: biomarker, cell and population phenotypes can be subtle, interconnected and system dependent. To deal with these challenges, we developed SimuCell (**Supplementary Software**; updated software available at http://www.SimuCell.org/), an open-source framework (**Fig. 1a**) for specifying and rendering realistic microscopy images containing diverse cell phenotypes, heterogeneous populations, microenvironmental dependencies and imaging artifacts.

SimuCell differs from existing cell-population generators[5] in three ways. First, SimuCell can generate heterogeneous cellular populations composed of diverse cell types. Each cell type can be defined independently by specifying models for cell and organelle shape and distributions of markers over these shapes. Models are typically algorithmic, but there is support for rendering produced by other tools, such as the highly realistic models learned from image data by CellOrganizer[3] (via the new SLML markup language). Second, SimuCell allows users to specify interdependencies among population, biomarker and cell phenotypes. For example, a marker's cellular distribution can be affected by the cell's microenvironment (**Fig. 1b**, marker 1) as well as the localization pattern of another marker (**Fig. 1b**, markers 2 and 3). These definable image properties are accessible to users either via a novel scripting syntax built on top of MATLAB or through a graphical user interface; intermediate results can be used to define further 'ground truths' (for example, cell boundaries can be used to validate segmentation algorithms). Finally, SimuCell is easily extensible, providing a standard framework for defining new plugins that can also be shared through the SimuCell website. Users interested in adding novel phenotypes to SimuCell's palette can typically do so by writing just a few lines of code, in part because of MATLAB's extensive library of functions. We also intend to implement a user forum to share ideas, scripts, plugins and images. Thus SimuCell allows the definition of a broad