IMAGE CONTENT-BASED RETRIEVAL AND AUTOMATED INTERPRETATION OF FLUORESCENCE MICROSCOPE IMAGES VIA THE PROTEIN SUBCELLULAR LOCATION IMAGE DATABASE

Kai Huang, Jennifer Lin, James A. Gajnak and Robert F. Murphy

Department of Biological Sciences and Center for Automated Learning and Discovery, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh PA 15213, USA, murphy@cmu.edu

ABSTRACT

We describe the Protein Subcellular Location Image Database (PSLID), which collects and structures 2-D through 5-D fluorescence microscope images, annotations, and derived features in a relational schema. It is designed so that interpretations as well as annotations can be queried. The annotations in PSLID, composed of 28 linked tables with publicly available descriptions, provide a thorough description of sample preparation and fluorescence microscope imaging. Image interpretation is achieved using Subcellular Location Features that have been shown to be capable of recognizing all major subcellular structures and of resolving patterns that cannot be distinguished by eye. Results of queries can be ranked by image typicality and statistical tests can be performed on sets of images (e.g., to determine whether a drug alters the distribution of a tagged protein). We anticipate that PSLID will serve as a common repository for microscopy images documenting the subcellular location of proteins.

1. INTRODUCTION

Advances in fluorescent probe chemistry and imaging techniques have made fluorescence microscopy a crucial method for cell and molecular biology [1]. However, most biologists still maintain their images without any metadata structure or quantitative description. Although some online biological image database systems [1-3] have tried to organize biological images from different sources by using a metadata structure to facilitate image exchange and management, the absence of rich numerical features for describing images limits the utility of these databases. Systems that provide more expanded searching of image collections have been described previously (e.g., the QBIC system, [4,5]), but these have used limited sets of general purpose image features and have not been applied to microscope images.

Our group has developed sets of Subcellular Location Features (SLF) and demonstrated that these features are sufficient to enable the recognition of all major subcellular structures [6,7]. In this paper, we describe the use of these features to allow a fluorescence microscope database system to perform queries by image content as well as via text annotations. Our system utilizes a publicly available database schema that provides searchable, thorough annotation of sample preparation and image collection. More significantly, we demonstrate that the use of the SLF can provide automated interpretation of the results of database searches.

2. ORGANIZATION OF PSLID

PSLID is designed as a relational database on the Oracle 8.1.7 platform. As shown in Table 1, each image in PSLID is accompanied by a description of the source of the image (e.g., researcher, laboratory, cell type), an explanation of the experimental details (e.g., protocol, label, dye, microscope, objective, illumination, filters), and a set of numerical features. The web-based process for loading images into PSLID provides a simple, controlled-vocabulary means for entering the annotations and also triggers calculation of the SLF features for each cell in each image.

The SLF features we have described consist of several different types, including Zernike moment features, Haralick texture features, and features derived from morphological and geometric image processing. A complete description of the individual features and the predefined feature sets is available at http://murphylab.web.cmu.edu/services/SLF/.

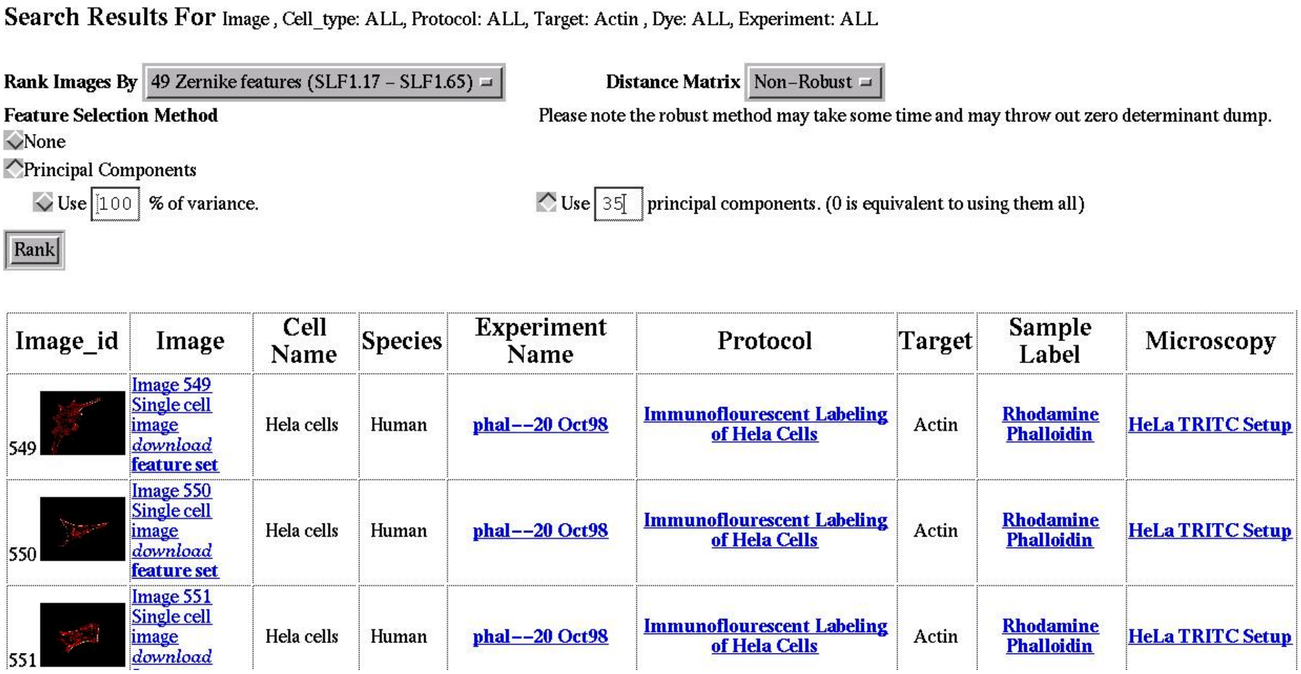
Some features (e.g., those in the Double_Image_Feature_Values table) may require a parallel, reference image (such as a DNA distribution). The sample description information in the database can be used to determine whether these features can be calculated. All features are given a standard nomenclature with a prefix SLF followed by the set number as well as an index.

Materialized views are created to speed keyword searches. Several secondary indexes are available in

0-7803-7584-X/02/$17.00 ©2002 IEEE

Table	Attributes	Description of contents
Laboratory	Laboratory_id, Name, Address, Email_address, Phone_number	Information about the laboratory conducting the experiment including contact information.
Researcher	Researcher_id, Laboratory_id, Name, Email_address, Phone_number	Information about the specific researcher conducting the experiment including contact information.
Experiment	Experiment_id, Researcher_id, Title, Date	General information about the experiment such as title and date started.
Slide	Slide_id, Cell_type_id Experiment_id, Protocol_id, Sample_label_id	Holds links to other tables. One record is created for each individual slide.
Protocol	Protocol_id, Title, Author, Reference, Fixation, Permeabilization, Substrate, Temperature, Protocol_text	Specific information about the protocol used to prepare the slides as well as a full text write-up of the procedure.
Cell_Type	Cell_type_id, Name, ATCC_Number, MESH_Heading, MESH_Tree_Number, Organism	Information on the type of the cell in the sample.
Sample_Label	Sample_label_id, Name	An entry for each unique combination of labels (probes).
Label_JUNC	Sample_label_id, Label_id, Position	Order relative to other labels in a labeling protocol.
Target	Target_id, Name, MESH_Heading, MESH_Tree_Number	Information on the macromolecule that is the target of the labeling.
Label	Label_id, Dye_id, Name, Target_id, Ligand_class	Information about individual labels and the macromolecule they bind to (e.g., IgG2a directed against the N-terminus of giantin).
Dye	Dye_id, Name, MESH_Heading Excitation_max, Emission_max	Detailed information about the dyes attached to the labels.
Field_of _View	Field_of_view_id, Slide_id X_coor, Y_coor	Information about a specific field on the slide.
Time	Time_id, Field_of_view_id	Assigns a unique id to all images taken at the same time.
Stack	Stack_id, Microscopy_id, Time_id Description, Target_id	Information about an entire stack of images. The target id may refer to a more general target than the target id in the Label table.
Image	Image_id, Stack_id, URL, Stack_pos, X_dim, Y_dim, Z_coor	Information on an individual 2D image.
Feature	Feature_id, Name, Description Num_dimensions, Num_channels	Characteristics of each numerical feature.
Single_Stack _Feature_Values	Single_stack_feature_values_id, Stack_id, Feature_id, Value	The features calculated from an individual 3D image.
Double_Stack _Feature_Values	Double_stack_feature_values_id, Stack_id1, Stack_id2, Feature_id, Value	The features calculated from two 3D images.
Single_Image _Feature_Values	Single_image_feature_values_id, Image_id, Feature_id, Value	The features calculated from an individual 2D image.
Double_Image _Feature_Values	Double_image_feature_values_id, Image_id1, Image_id2, Feature_id, Value	The features calculated from two 2D images.
Microscopy	Microscopy_id, Name, Objective_id, Detector_id, Condenser_id, Illumination_id, Microscope_id, Emission_magnification, Exposure_time	Information about microscopy set up.
Detector	Detector_id, Name, Type, Manufacturer, Model, S/N, Gain, Voltage, Offset, X_res, Y_res, Bit_depth, Detector_processing	Information about detector set up.
Microscope	Microscope_id, Name, Manufacturer, Model, S/N	Information about each microscope.
Filter	Filter_id, Name, Pass_type, Upper, Lower, Thickness, Spectrum	Characteristics of an individual filter.
Microscopy _Filter	Microscopy_filter_id, Microscopy_id, Filter_id (Ex), Excitation_filter_angle, Filter_id (Em), Emission_filter_angle	An entry for each unique combination of filters.
Objective	Objective_id, Name, Manufacturer, Model, S/N, Magnification, Num_app	Information about each objective.
Condenser	Condenser_id, Name, Manuf., Model, S/N	Information about each condenser.
Illumination	Illumination_id, Name, Type, Manufacturer, Model, S/N, Spectrum	Information about each illumination source.

Table 1. Summary of Fluorescence Microscopy Annotation Schema, version 1. A complete description of each table and field is available at http://murphylab.web.cmu.edu/services/FMAS. S/N refers to serial number.



Search Results For Image , Cell_type: ALL, Protocol: ALL, Target: Actin , Dye: ALL, Experiment: ALL

Rank Images By 49 Zernike features (SLF1.17 – SLF1.65) **Distance Matrix** Non–Robust

Feature Selection Method

Please note the robust method may take some time and may throw out zero determinant dump.

◇None

◇Principal Components

◇ Use 100 % of variance. ◇ Use 35 principal components. (0 is equivalent to using them all)

Rank

Image_id	Image	Cell Name	Species	Experiment Name	Protocol	Target	Sample Label	Microscopy
549	Image 549 Single cell image download feature set	Hela cells	Human	phal--20 Oct98	Immunoflourescent Labeling of Hela Cells	Actin	Rhodamine Phalloidin	HeLa TRITC Setup
550	Image 550 Single cell image download feature set	Hela cells	Human	phal--20 Oct98	Immunoflourescent Labeling of Hela Cells	Actin	Rhodamine Phalloidin	HeLa TRITC Setup
551	Image 551 Single cell image download	Hela cells	Human	phal--20 Oct98	Immunoflourescent Labeling of Hela Cells	Actin	Rhodamine Phalloidin	HeLa TRITC Setup

Figure 1. A screen copy showing part of the results from choosing the target protein Actin (with no other restrictions). The top of the page allows specification of parameters to be used for ranking the images using TypIC.

every table as well as in the materialized views to accelerate searches and eliminate duplicates.

3. ACCESS TO PSLID

A web interface of the PSLID is implemented through Java Server Pages (JSP). The web interface permits querying on cell type, protocol type, protein target, dye type, experiment title, etc. After the user specifies the query and the maximum number of images to display, the JSP will communicate with the database through a JDBC interface.

The resulting page (Figure 1) contains information about the query, a thumbnail picture of each individual cell image and links to: display the full resolution cell image, display the features of that image, display the full resolution original image (which might contain more than one cell), download the original image, and to display several pages describing experiment setup, protocol, sample labeling, and microscopy setup.

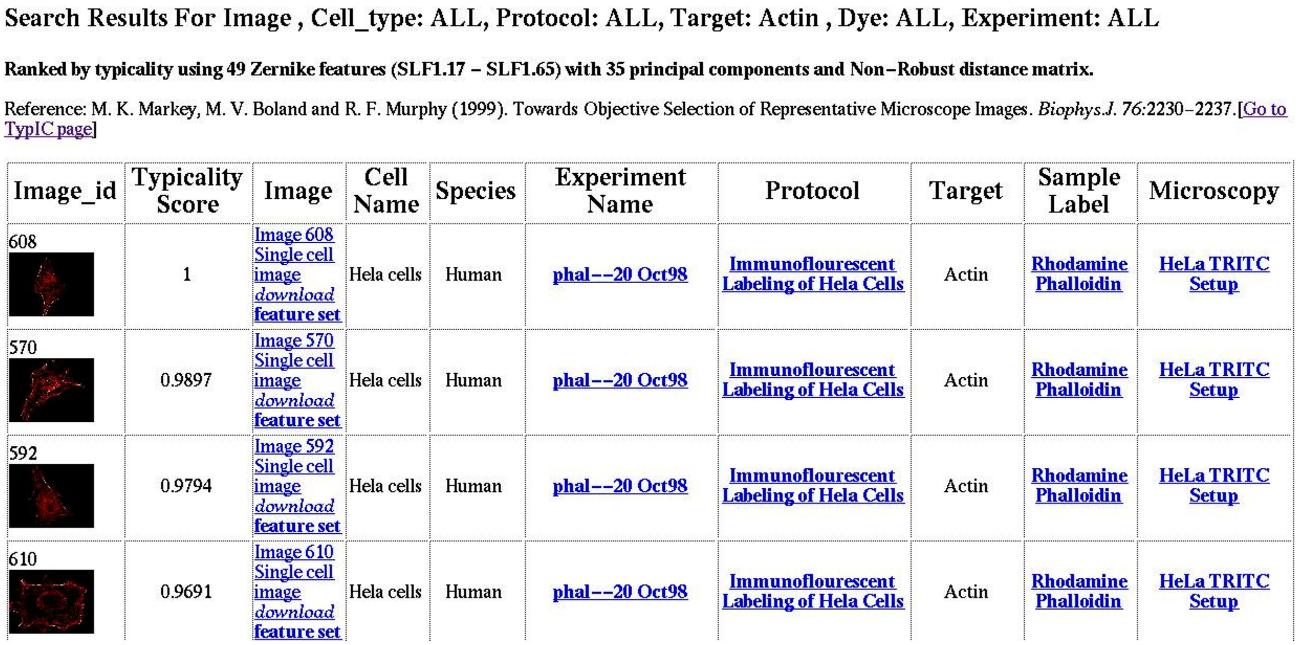
Due to the large number of images that may be returned from a specific query, we incorporate our previous method for typical image selection, TypIC [8], in the result page (Figure 1). The user is allowed to choose the feature set used for the ranking, whether principal components will be selected from the feature set, and whether outlier detection will be used (referred to as Robust covariance matrix estimation). This application can guide users to the most typical images in the query results so that they will have a better conception of the underlying pattern and may download only the most useful images for their purpose (Figure 2).

A common use of microscopy is to determine whether the distribution of a particular protein is changed by a particular experimental manipulation (e.g., the addition of a drug). We have therefore incorporated our SImEC service [9] into PSLID to compare two result sets saved in the server. This function can help the user find potential similarity between two queries or to determine whether a change in conditions has resulted in a statistically significant change in a protein pattern.

A final service provided through PSLID is the classification of returned images. Back-Propagation Neural Networks and Support Vector Machines that have been trained on all location classes are available to classify images returned from queries. Currently, our classifiers require that each image contain only a single cell (i.e., have been cropped to remove other cells prior to inclusion in the database).

4. CONCLUSIONS

Our group is the first to carry out systematic studies of protein subcellular location via fluorescence microscopy and to provide validated numerical descriptors so that protein patterns can be compared and classified. The capabilities derived from our previous work have been incorporated into PSLID to provide a comprehensive application incorporating a relational database, machine



Search Results For Image , Cell_type: ALL, Protocol: ALL, Target: Actin , Dye: ALL, Experiment: ALL

Ranked by typicality using 49 Zernike features (SLF1.17 – SLF1.65) with 35 principal components and Non–Robust distance matrix.

Reference: M. K. Markey, M. V. Boland and R. F. Murphy (1999). Towards Objective Selection of Representative Microscope Images. *Biophys.J. 76:*2230–2237.[Go to TypIC page]

Image_id	Typicality Score	Image	Cell Name	Species	Experiment Name	Protocol	Target	Sample Label	Microscopy
608	1	Image 608 Single cell image *download* **feature set**	Hela cells	Human	**phal––20 Oct98**	**Immunoflourescent Labeling of Hela Cells**	Actin	**Rhodamine Phalloidin**	**HeLa TRITC Setup**
570	0.9897	Image 570 Single cell image *download* **feature set**	Hela cells	Human	**phal––20 Oct98**	**Immunoflourescent Labeling of Hela Cells**	Actin	**Rhodamine Phalloidin**	**HeLa TRITC Setup**
592	0.9794	Image 592 Single cell image *download* **feature set**	Hela cells	Human	**phal––20 Oct98**	**Immunoflourescent Labeling of Hela Cells**	Actin	**Rhodamine Phalloidin**	**HeLa TRITC Setup**
610	0.9691	Image 610 Single cell image *download* **feature set**	Hela cells	Human	**phal––20 Oct98**	**Immunoflourescent Labeling of Hela Cells**	Actin	**Rhodamine Phalloidin**	**HeLa TRITC Setup**

Figure 2. A screen copy showing the results from the query in Figure 1 after ranking with TypIC.

learning, and statistical inference. It is an example of applying data mining on top of a relational database and achieving query interpretation. The annotation schema and interpretation capabilities of PSLID will be merged with the nascent OME (Open Microscopy Environment, http://openmicroscopy.com) project to provide increased accessibility of the methods we have developed to the fluorescence microscopy community. We anticipate that PSLID through OME will serve as a common repository for microscopy images documenting the subcellular location of proteins.

ACKNOWLEDGMENTS

This work was supported in part by NIH grant R33 CA83219 and by NSF grant MCB-8920118. K.H. and J.L. were supported by a Graduate Fellowship and a Summer Scholar Award, respectively, from the Merck Computational Biology and Chemistry Program through a grant from the Merck Company Foundation. We thank Dr. Michael Boland for generating an initial draft of the database schema, and Meel Velliste, Edward Roques and Mia Markey for helpful discussions.

REFERENCES

[1] P. Andrews, I. Harper, and J. Swedlow, "To 5D and Beyond: Quantitative Fluorescence Microscopy in the Postgenomic Era," *Traffic* 3, 29-36, 2002.

[2] J. M. Carazo, and E. H. K. Stelzer, "The BioImage database project: Organizing multidimensional biological images in an object-relational database.," *Journal of Structural Biology* 126, 97-102, 1999.

[3] E. Gonzalez-Couto, B. Hayes, and A. Danckaert, "The life sciences Global Image Database (GID).," *Nucleic Acids Research* 29, 336-339, 2001.

[4] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, H. Qian, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: the QBIC system," *Computer* 28, 23-32, 1995.

[5] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin, "The QBIC project: Quering images by content using color, texture, and shape," *Proceedings of the SPIE Conference on Storage and Retrieval for Image and Video Databases* (pp. 173-187), 1993.

[6] M. V. Boland, M. K. Markey, and R. F. Murphy, "Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images," *Cytometry* 33, 366-375, 1998.

[7] M. V. Boland, and R. F. Murphy, "A Neural Network Classifier Capable of Recognizing the Patterns of all Major Subcellular Structures in Fluorescence Microscope Images of HeLa Cells," *Bioinformatics* 17, 1213-1223, 2001.

[8] M. K. Markey, M. V. Boland, and R. F. Murphy, "Towards objective selection of representative microscope images," *Biophysical Journal* 76, 2230-2237, 1999.

[9] E. J. S. Roques, and R. F. Murphy, "Objective Evaluation of Differences in Protein Subellular Distribution," *Traffic* 3, 61-65, 2002.