CellProfiler Analyst™ data exploration software

Created in the laboratories of Anne E. Carpenter, David M. Sabatini and Polina Golland at:

And now based at:

Authors (in order of joining the project):
In Han Kang
Adam Papallo
Adam Fraser
INTRODUCTION ............................................................................................................... 4
INSTALLATION .............................................................................................................. 5
  ACCESSING A MySQL OR Oracle SERVER .......................................................... 5
  SETTING UP THE EXAMPLE DATABASE ................................................................ 5
  SYSTEM REQUIREMENTS ......................................................................................... 5
  CELLPROFILER ANALYST INSTALLATION GUIDE (ALL PLATFORMS) ....................... 6
SETTING UP A PROPERTIES FILE ............................................................................. 7
  DATABASE TYPE .................................................................................................. 8
  DATABASE HOST NAME/IP ADDRESS ........................................................................ 8
  DATABASE PORT .................................................................................................... 8
  DATABASE USERNAME/PASSWORD .......................................................................... 8
  DATABASE NAME .................................................................................................. 8
  IMAGE/Object TABLES ........................................................................................... 8
  UNIQUE IMAGE IDENTIFIER .................................................................................. 8
  IMAGE PRIMARY OBJECT COUNT COLUMN .......................................................... 9
  UNIQUE OBJECT IDENTIFIER IN IMAGE ............................................................... 9
  INFORMATION TABLE ............................................................................................. 9
  TREATMENT INFORMATION COLUMN ................................................................ 9
  WEB INFORMATION COLUMN ............................................................................... 9
  WEB INFORMATION URL PREPEND ................................................................. 9
  IMAGE TRANSFER PROTOCOL / IMAGE ACCESS PREPEND .................................. 9
  IMAGE FORMAT INFORMATION ........................................................................... 10
  IMAGE PATHWAYS AND FILENAMES ................................................................... 10
  X/Y COORDINATES FOR PRIMARY OBJECT CENTER ......................................... 10
  CLASSIFY BY GROUP TABLE / CLASSIFY BY GROUP COLUMN(s) ....................... 10
  WIDTH OF OBJECT CROPPING SQUARE .............................................................. 11
FILE MENU .................................................................................................................. 11
  NEW PROPERTIES ................................................................................................ 11
  EDIT CURRENT PROPERTIES .................................................................................. 11
  SAVE PROPERTIES / SAVE PROPERTIES AS ........................................................ 11
  MAKE LOCAL OBJECT TABLE .............................................................................. 11
  LOAD TAB-DELIMITED TABLE .............................................................................. 12
DISPLAY MENU ......................................................................................................... 12
  GENERAL PLOTTING INTERFACES ...................................................................... 12
  SCATTERPLOT ......................................................................................................... 14
  DENSITY PLOTS ...................................................................................................... 15
  HISTOGRAM ........................................................................................................... 16
  PARALLEL COORDINATES ................................................................................... 17
  TABLE/IMAGE ......................................................................................................... 17
  GATING .................................................................................................................... 18
ANALYSIS MENU: CLASSIFIER ................................................................................ 19
  RETRIEVE OBJECT IMAGES .................................................................................. 19
  CLASSIFY OBJECT IMAGES TO CREATE A TRAINING SET .................................. 20
  TRAINING CLASSIFIER (FIND RULES) ................................................................. 20
  REPEAT TRAINING ITERATIVELY ......................................................................... 21
  SCORE IMAGES ...................................................................................................... 22
  SAVE/LOAD TRAINING SETS ............................................................................... 22
  SAVE/LOAD/ADD RULES ...................................................................................... 22
Introduction

CellProfiler Analyst software allows interactive exploration and analysis of multidimensional data, particularly data from high-throughput, image-based experiments. It also contains a supervised machine learning system that can be trained to recognize complicated and subtle phenotypes, enabling automatic scoring of millions of cells.

Measurements can be explored on an individual-cell basis, or grouped by image or sample condition. Interesting samples can further be explored by linking to the original image that produced the numerical data or to web-based information about the sample. A subset of data points in one plot can be selected and used to generate a new plot focused on only that group, or can be highlighted in all displayed plots to explore relationships in the data.

Compatible with MySQL and Oracle databases (note that Oracle support is currently in progress but testing as yet is incomplete), CellProfiler Analyst is especially suited to explore the huge datasets produced by CellProfiler open-source cell image analysis software, where a typical experiment contains ~1000 image-based measurements for each of ~100 million cells.

The project is based at the Broad Institute Imaging Platform and was originally founded by InHan Kang, Thouis (Ray) Jones, and Anne E. Carpenter in the laboratories of David M. Sabatini at the Whitehead Institute for Biomedical Research and Polina Golland at the MIT Computer Science and Artificial Intelligence Laboratory.
Installation

Download the latest version of CellProfiler Analyst at http://www.cellprofiler.org

Accessing a MySQL or Oracle server
Before using CellProfiler Analyst, you will need to be able to access a MySQL or Oracle server (Note that we have tested CellProfiler Analyst with MySQL extensively but not with Oracle). For questions on installing, configuring, and running a DB server, please refer to http://mysql.com, http://oracle.com, or contact your IT department; no further instruction is provided here.

Setting up the example database
An example MySQL database is available for download at www.cellprofiler.org/examples. It is best to have an experienced DBA set up this database; only minimal instructions are provided here. Briefly, to set up the example database to be accessed by CellProfiler Analyst:
1. If your installation of MYSQL is not running, execute this command (assuming mysql is installed in /usr/local):
   `sudo /usr/local/mysql/bin/mysqld_safe`
2. Load the data into MySQL. Adjust the username depending on the MySQL configuration; for example, in the default install, "DBUserName" should be changed to "root":
   `/usr/local/mysql/bin/mysql -u DBUserName < exampleSETUP.SQL`
3. Move the "images" directory to the same location as you have installed the CellProfiler Analyst Jar file (CPAnalyst.jar).
4. Launch CellProfiler Analyst
5. Load the "example.properties" (Under "File -> Open Properties"). It may be necessary to adjust the settings in the property file to match the accounts and database information, depending on your MySQL installation.

System Requirements
1. 512 MB of RAM or greater
2. OS capable of running Java applications (e.g., OS X v 10.4, Windows XP, Linux or other).
3. Java 1.4.2 or greater (Java 1.5.0_6 or greater is required for full functionality). The latest version can be downloaded from http://java.sun.com/. You can check your version by typing `java –version` at the command prompt:
   o Mac: the command prompt is accessed by /Applications/Utilities/Terminal
   o Windows: the command prompt is accessed by Programs -> Accessories -> Command Prompt or Programs -> MS-DOS Prompt
4. For full functionality, you will need Python version 2.5 or greater (can be downloaded from http://www.python.org/) and The NumPy Python package (can be downloaded from http://scipy.org).
1. Extract the contents of the CPAnalyst.zip file to any directory (the directory cannot have any spaces in its name).
2. CPAnalyst should be started using CPAnalyst.command (on Mac or Unix) or CPAnalyst.bat (on Windows).
3. For convenience, you can add a shortcut to the CPAnalyst.command file (by dragging its icon to the dock bar in OS X, or using the window manager in Unix) or the CPAnalyst.bat file (adding a shortcut to the desktop) to make launching CellProfiler Analyst easier.
4. If you are using the example MySQL database, the accompanying “images” directory should be placed in the CPAnalyst directory created in step (i). It may also be necessary to adjust the username, password, and database information in the example.properties file.
Setting Up a Properties File

After starting CellProfiler Analyst (by clicking on the CPAnalyst.term, or CPAnalyst.bat, or CPAnalyst.command, file or shortcut), you would normally load a Properties File. Properties files supply CellProfiler Analyst with the information necessary to access your data. You can create a new set of properties by choosing New Properties from the File menu. A form like the one shown in Figure 1 will be displayed.

Figure 1: Property File Configuration Window

To modify these properties correctly you will need to know some information about where and how your data is stored. If you do not know any of the below information you should contact your database administrator (DBA) for assistance. Once you have completed entering your properties, click “OK.”
If you discover that you made a mistake, choose Edit Current Properties from the File menus to make modifications. Once you have a working set of properties, choose Save Properties from the File menu to save them to a file.

The properties are as follows:

**Database Type**
Database type refers to the RDMS (relation database management system) in which your data is being stored (e.g. MySQL or Oracle).

**Database Host Name/IP Address**
The machine that is storing your data in the RDMS is referred to as the database host. This machine will be accessed either through its host name (e.g. localhost or dbServerName.domain) or IP address (e.g. 127.0.0.1).

**Database Port**
In order to communicate the information from one computer to another, an interface or port is used. A particular RDMS has its own port through which it communicates data. This port has a unique identifier. In the case of MySQL the port used is identified by 3306 and in Oracle 1521, unless set otherwise by your DBA.

**Database Username/Password**
Your username and password grants you access to the RDMS and the schema to which you have permissions.

**Database Name**
The database name is the schema that you wish to access with this properties file. Most likely this is the name of the schema for a specific experiment. If you created the database using CellProfiler, this is the text you entered for “Database Name” in the Export to Database module.

**Image/Object Tables**
Image Table(s) and Object Table(s) are fields containing lists of the tables that contain the quantitative data that you wish to access with this properties file, separated by commas in their respective field (e.g. Image Tables: expA_per_image, expB_per_image and Object Tables: expA_per_object, expB_per_object. Note that CellProfiler ExportToDatabase default is “Per_Image” and “Per_Object”). An image table contains statistics from the image as a whole (e.g. the number of cells per image, or the mean intensity of nuclei in the image), while an object table contains information on the individual objects (e.g. cells, worms, or yeast colonies) that were identified in your images.

**Unique Image Identifier**
The unique image identifier is the column that identifies uniquely the row of data that pertains to a given image. This is “ImageNumber” for CellProfiler output.
**Image Primary Object Count Column**
In the image table(s) outputted by CellProfiler there exists a column that contains the counts for your biological objects of interest, this column is referenced here.

**Unique Object Identifier in Image**
Objects in the object table(s) are uniquely identified by the image from which they came and an object number. The latter of these is the unique object identifier in the image. For CellProfiler data this is “ObjectNumber”.

**Information Table**
The Information Table contains the annotation for each image that serves to link the quantitative data to experimental treatment. If such a table does not exist this field may be left blank. For example, this table might contain chemical compounds, chemical concentrations, or gene names, depending on the type of experiment that was performed.

**Linker Column for Information and Image Tables**
The column that is shared by both the image table(s) and information table links the quantitative data to the annotation data.

**Treatment Information Column**
A treatment information column for a particular experiment encapsulates a particular human readable column of annotation in the information table (e.g. a column containing gene names or compound treatments). This column should be selected from the Information Table.

**Web Information Column**
In the information table, this is a column that contains an identifier that will link images to external web databases such as NCBI, FlyBase, ChemBank, or PubMed.

**Web Information URL Prepend**
The URL that will come before the identifier whose entry on that website you wish to access. Two common web information URL prepends are:
- http://flybase.bio.indiana.edu/_.bin/fbidq.html? followed by a FlyBase identifier

**Image Transfer Protocol/Image Access Prepend**
An image transfer protocol (ITP) is the method by which images will be communicated to CellProfiler Analyst. Supported ITPs are http, local, smb, and ssh. Some supported transfer protocols require use of the Image Access Prepend (IAP). The IAP is put before the path to an image such that CellProfiler Analyst is granted permission to view the image. The following rules will help you choose the appropriate ITP and IAP:

1. Use http if your images are located on a web server
   - Set the IAP to the necessary web prefix for the image path/filename (e.g. http://link_to_your_images.edu/)
ii) Use local if your images are on your local machine (or accessible through a mounted drive)
   a. IAP should be left blank
iii) Use smb if your images are located on an smb server
    a. Set the IAP to access the smb server with your username,
       password, and relevant server information (e.g. smb://username:password@server_name)
iv) Use ssh if your images are located on a machine, other than your local
    machine, that is not a server but has ssh capabilities.
    a. Set the IAP to a comma separated list containing your username,
       computer name or IP address, and password. (e.g. username,IP,password)

**Image Format Information**
This is where you specify what type of images you have.
   i) For images in the DIB format (assumed to be square images), supply the
      width of the images in pixels.
   ii) For images in the TIFF format that are 12 bit encoded as 16 bit, type Y here.
   iii) For images in other formats (regular TIFF, jpeg, or bmp), type N here.

**Image Pathways and Filenames**
The image pathway and filename fields allow you to specify the color you would
like a given image (or channel) to take. For example if you have images
containing an actin channel and would like to display this as red, type the image
table column name that corresponds to the pathway for the actin channel images
in the Pathway to Red Images field and the image table column name that
contains the filename for the actin channel images in the Filenames for Red
Images field. (Note: If you have only two channels then you must use the red
and green fields first; moreover, if your images are color or you only have a
single channel then only enter data for the red pathway and filename fields.)

**X/Y Coordinates for Primary Object Center**
Indicate the columns that contain the x and y coordinates for the centers of the
objects that were used as the primary objects in the segmentation of the images.

**Columns to Ignore when Classifying**
When classifying objects you do not want to classify on columns that contain
arbitrary information such as object table key or other non-phenotypic measures.
This field allows you to specify which columns you want to exclude from use in
classification.

**Classify by Group Table / Classify by Group Column(s)**
When classifying you may want to score your images on the basis of groups from
common annotation information or metadata; you would supply the table name
containing this information in the Classify by group table. (Note: If the Classify
by group table is the same as the information table type: info_table.) In the
Classify by Group Column(s) you would supply the columns in the Classify by
Group Table by which you wish to group.
**Width of Object Cropping Square**

Type the width for a square-cropping window to be drawn around the center of individual object. This cropped image is what will be used in the display of object montages.

**File Menu**

The File menu (Figure 2) contains the options that allow you to specify the data that you want to access.

![File Menu](image)

**Figure 2: File Menu**

**New Properties**

Allows you to specify a new set of properties. The new properties are not automatically saved (see Save Properties below). The fields are described in the previous section, **Setting Up a Properties File**.

**Open Properties**

Allows you to open a set of properties that have previously been saved to a file.

**Edit Current Properties**

Allows you to make modifications to the current set of properties. The new properties are not automatically saved (see Save Properties below). Note that changing some fields of the properties file will require that you restart the application before changes take effect. The fields are described in the previous section, **Setting Up a Properties File**.

**Save Properties/Save Properties As**

These options save the current properties to a file. Save Properties As asks for a new file name.

**Make Local Object Table**

Selecting this generates a window for you to enter a comma-separated list of object table columns in the format `objectTable.column`. The purpose of this
option is to make accessing data quicker. It should be noted however, that this is not necessary when the connection speed to the server is sufficiently fast.

**Load Tab-Delimited Table**
This is an option that will prompt you for a file containing a tab-delimited table. Though it is not necessary for the file to contain web information and image columns, the interface prompts the you to enter zero indexed column values for columns containing this information. This is useful for uploading results from a previous CellProfiler Analyst session.

**Display Menu**
The Display menu (Figure 3) contains the options that are available to visualize and explore the data in the selected database.

![Figure 3: Display Menu](image)

**General Plotting Interfaces**
When you select any plot type from the Display menu, you will be prompted with a create plot window. All CellProfiler Analyst plots, with the exception of parallel coordinates whose table tab differs in design, have the following options in common:

*Table tab* (Figure 4): This tab allows you to set the type of data to plot. First, you choose whether each data point in the plot should represent an individual object or a whole image (this latter case includes when whole image measures are derived from the population of individual object measures in that image). Next, use the drop down menus to specify the actual measure to plot and on which axis. Finally, choose the axes scale: either linear or logarithmic (with a base of ten). You may wish to choose the logarithmic scale when plotting intensity measures or other measures where the underlying physics suggests that the measures will be distributed as function of an exponentiated random variable. (Note that in density plots both axes are either linear or logarithmic. That is to say you cannot have the X axis in linear scale and the Y axis in logarithmic scale for this type of plot.)
Options tab (Figure 5): This tab allows you to limit or group the data that you selected to plot in the Table tab. That is, you can limit your analysis to a single specified image (identified by its unique identifier) whose data you wish to plot, to a sample of data-points at specified equal intervals, or to data that satisfies criterion specified in SQL “where” clauses (excluding the WHERE keyword (e.g. ImageNumber > 100, could be an entry in this dialog box)). A final option presented by the Options tab allows you to group the data using the drop down menus to select the desired column names in the image or information tables.

Figure 4: The Create Plot window, with the Table tab selected, allowing you to select the data to be plotted.
Context menu within plots: Right-clicking on a plot provides another common interface for the data. With the exception of parallel coordinates, all of the plots permit some form of data manipulation via the right click. Plot properties such as figure title, axis label, text and background colors can be modified by using right-click: Properties. Other right-click features include entering select mode for gating off populations, as well as options related to selecting points such as highlighting and sub-plotting to name a few. Note that both the select mode and select related options are options that can also be accessed through menu bar selections.

Saving plots can be accomplished by selecting Save Image of Plot from the plot’s file menu.

Scatterplot
A scatterplot is simply a method of comparing two sets of numerical values in the specified database; one measure is plotted on the X axis and the other measure is plotted on the Y axis (Figure 6).

After making selections in the Table and Options tabs, click OK and the scatterplot will be displayed. Data points are red and +/- two standard deviations from the mean are represented as blue lines. Gating is especially useful for exploring scatterplots (see later section on gating).
Density plots

A density plot (Figure 7) represents the density of points in a plane, overcoming a limitation of scatterplots: when a scatterplot becomes too crowded, it becomes impossible to see how many points are in some regions of the plot. Density plots overcome this problem by displaying the density of data points as color. In CellProfiler Analyst, this density is represented by the color scheme that is found on the right of the density plot. The density plot is particularly useful when there are more than 10,000 data points to be plotted. To plot a density plot in CellProfiler Analyst, simply select the Density Plot option from the Display menu and select the necessary information from the Table and Options tabs. The density plot also has its own tab that allows you to trim off the upper and lower 0.1% of the data, thereby limiting the effect of outliers on the plot.
As with the scatterplot, you can select points of interest in the density plot. You do this by clicking the select button at the bottom of plot, gating the region of interest using a click-drag-release movement over the region and then selecting Highlight Gated under the Select Points menu. Once you have highlighted the gated points, you have a number of options available to explore the data in that region, most of which are self-explanatory. Compute Enrichments calculates how enriched or un-enriched each image is for the gated objects, relative to either the entire population or only the displayed population (in the event that the given density plot is displaying only a subset of all objects, that is, it is a sub-plot based on another plot). Figure 7 shows examples of the density plot and several of these options.

Figure 7: A density plot and results from using the Show Object Montage, Show Image of Random Point, and Compute Enrichments options

Histogram
The histogram function in CellProfiler Analyst allows you to select a single measure to plot as a histogram. As with other CellProfiler Analyst plots, you can choose a logarithmic (base 10) or linear axis and limit your search to a specific image, sample of images, or grouping. There is also a specific histogram tab when initializing the histogram to plot the data. This tab allows you to specify how many bins you want and on what range. If you do not specify values in the histogram tab, CellProfiler Analyst will default to fifty bins and have the range span the entire set of values for the chosen measure. Example histograms are shown in Figure 9.
Parallel Coordinates

Parallel Coordinates is a plotting method whereby you select multiple measures to form the x-axis and their scaled (0-1) values to form the y-axis (Figure 8). The interface for parallel coordinates has you select a data type and table from which to choose the measures to be plotted just as in the other interfaces. The difference between the parallel coordinates’ interface and the others is that the “Add” button must be clicked after each measure is selected. The Options tab has the same functionality as in other plotting interfaces.

Upon plotting the data, you will notice that there are only integers for axis labels on the plot; this is to limit clutter at the bottom of the plot as many of the measure names are quite long and therefore take up more space than is allotted. To view the measure names simply use the Show Info menu at the top of the plot frame and select Show Axes. This brings up a table with the axis key.

One other key difference of parallel coordinates from all other plots is that in order to select images one must use the Select Points menu and click the Select Region option. This will allow you to click and drag your mouse over a region of interest. The other options for parallel coordinates are self-explanatory.

Figure 8: Example of Parallel Coordinates plotting

Table/Image

The Display Table and Image options work as their names suggest. Table allows you to select a set of columns of data from a table and view it all or a subpopulation that can be defined in the Options tab. Image allows you to select an image or an image with one of the objects highlighted, should you choose the object table as the data set to query.
Gating

Gating is a convenient tool that allows you to explore scatterplots, and some other plots, in CellProfiler Analyst (Figure 9). You can gate off part of the current plot to find out more information about the data points contained in that region of the plot by using the Show Info and Select Points menu bars. For example if you have created a scatterplot of whole-image data and are interested in those images with a mean intensity of DNA more than two standard deviations above the mean of all images and an integrated actin intensity more than two standard deviations below the mean, you would click the Select button located at the bottom of the plot, highlight that region of the plot, and then go to Select Points in the tool bar and select Highlight Gated option. This will change the color of the points within the gate from red to blue as well as allow methods found under Show Info and Select Points to be performed on these points. One such option is the Plot Selected in New Scatterplot option found under the Select Points menu to plot just those points in a new plot. In the new plot you can change the axes by clicking on them to get a new plot just for these points. Then by looking at histograms for the two axes of your latest plot you can determine where in the distribution the selected points lie. If you are interested in viewing a particular image or set of images represented by these points you can highlight them and using the Show Info drop down menu option, Show Image, and CellProfiler Analyst will display the selected images. Alternatively you can select a random image from the set of highlighted points (Figure 9).

Figure 9: Example plots demonstrating gating
Analysis Menu: Classifier

Choosing Analysis > Classifier displays the Classifier window (Figure 10). These tools allow you to train CellProfiler Analyst to identify objects of interest, by machine learning methods applied to the measurements for each object in your database. The machine learning algorithm uses the training set of positive and negative objects that you develop to determine a set of rules that best distinguishes the positive and negative cells in the measured feature space.

![Classifier window](image)

**Figure 10: Classifier window**

Retrieve object images

You should begin classification with the top panel titled “Grab Cells to Classify”. There, you can retrieve images of objects to begin training. Do this by changing the text fields and combo boxes in the retrieval phrase, and then click the “Grab Cells” button. Note that the text and controls in this panel form a “retrieval phrase.” We detail the retrieval process by describing each of the controls laid out in the retrieval phrase from left to right:

- The first box specifies the number of cells you want to look at. This is defaulted to 24, but you can enter any positive number or “all” to grab all cells from a particular image.
- The second control in the retrieval phrase is a combo box that specifies the type of cells to grab. Use random if you have not yet trained the classifier. Once a set of rules has been created, you can choose to search the database for cells that test positive or negative against the current rule set.
- The third control is another combo box that simply specifies whether you would like to see cells from a particular (“specific”) image only (faster), or “any” image in the current table (slower).
- Fourth is a text field where you can specify the image number to retrieve from. Note, if you select “any” in the previous control, this box is irrelevant and is hidden to preserve the readability of the retrieval phrase.
At the end of the retrieval phrase is a combo box that lets you choose the table from which you would like to retrieve cells.
Finally, the “Grab Cells” button performs the cell retrieval.

Classify object images to create a training set

Once retrieved, the cells are displayed in the “Unclassified Cells” bin in the bottom panel of the window. Drag and drop each cell in the Positive or Negative bin. If you are unsure whether to count an object as positive or negative, you may leave it in the Unclassified bin and it will be ignored and replaced the next time you retrieve objects. Multiple cells can be selected at a time by holding shift while clicking on the cells. To move a group of selected cells, right-click and select “Move to [Bin].” You can also select all cells in a board by right-clicking and choosing select all.

You can change the colors displayed by using the “Colors” menu at the top of the window. Also, if you would like to see the whole image that a particular selected cell(s) came from, you can do so by making a selection, then right-clicking and selecting “Show Full Images of Selected Cells.” This can be useful if a particular well/image contains many cells exhibiting a rare phenotype you are trying to classify. Check the window title for which image (“spot”) number the cell came from, and then grab more cells from that image. The size of the window in which each cell is cropped can also be changed (see Setting Up a Properties File: Width of Object Cropping Square).

The collection of positive and negative example objects is called a training set. You should classify many objects into the positive and negative bins (that is, create a reasonable sized training set) before attempting to train CellProfiler Analyst to identify positive objects. The number of objects to classify before beginning to train the computer is a critical parameter. Too few and the computer may learn an over-simplified definition of the objects that you are trying to classify. Too many and you have wasted your time classifying objects by hand. The optimal number depends very much on the complexity of the objects’ appearance and the scarcity of positive objects in the experiment. A good rule of thumb is to create a training set with at least 50 positive and negative objects before training. One can save a training set by selecting “Save Training Set” from the file menu, and load it again by selecting “Load Training Set.”

Training Classifier (Find rules)

Training is done in the Classifier window in the center pane titled “Train Classifier.” After classifying a suitable number of objects, click the Find Rules button to begin training CellProfiler Analyst to identify positive and negative objects. Clicking this button starts the classify function that examines what measures best discriminate between positive and negative objects in the training set and develops rules to carry out this automatic classification. After calculating, the Classifier will display the rules that CellProfiler Analyst found to distinguish the positive and negative objects.
You can change the maximum number of rules used for this training by typing in the “Max # of rules to find” text box. If the number of rules used by CellProfiler Analyst is less than the number of rules specified, it is because the positive and negative populations are distinguishable by a smaller number of rules. The number of rules used to train CellProfiler Analyst is a critical parameter. Too few rules and the accuracy of CellProfiler Analyst’s automatic classification will not be very high. Though, there is little risk of over-fitting with this machine learning method1, too many rules wastes calculation time, since this is an intensive step. The optimal number depends very much on the complexity of the objects’ appearance and the scarcity of positive objects in the experiment. A good rule of thumb is somewhere in the range of 10-50 rules. To find the optimal number of rules over a range of 0 to N (where N is the maximum number of rules that you wish to use in classification) use the X-Validate button, which will require significant computation time. This option calculates the truth table via k-folds cross validation algorithm at each value n in the range of 0 to N rules. From the truth table you can calculate which number of rules minimizes the number of false positives and false negatives by minimizing the square root of sum of the square of the false positive rate and the square of the false negative rate. The number of rules that minimizes that function should be the number of rules applied to the entire dataset. Note that cross-validation does not yield the classifier’s accuracy, exactly. See next section.

The rules themselves are the result of applying an implementation of gentle boosting to weak learners. This is an algorithm that sequentially finds the best rules that minimize the weighted least squares error when applied to the training set. A rule itself is a logical “if statement” that informs the computer what value to return on each object based on that object’s measured value for that feature. The value returned is either a positive or a negative number. These numbers correspond into the weighted partial least squares error as calculated when applied to the positively and negatively labeled objects of the training set. For example, if the rule returns a positive value for an object that means that by this rule the object falls into the set of positively labeled objects. By summing up all of the values returned by the rules for each object the computer classifies each object as either a positive or negative object based on the sign of the sum.

**Repeat training iteratively**

The rules allow CellProfiler Analyst to attempt automatic classification of positive and negative objects from the experiment. Depending on the complexity of the objects’ appearance, the rules usually do not allow very accurate classification after only one round of classifying objects into bins. Rather, iteration is an efficient means to improve the accuracy. After using Find Rules, you now have the option to grab positive or negative cells, in addition to random cells. These objects appear in the Unclassified bin, where you can drag them to their proper bin as before (Figure 11). In this iterative stage of training, it is most effective to drag misclassified objects to their proper bin (for example, if you clicked to retrieve positive objects and obtained a few negative objects among the

---

positives, take special care to drag the negatives to the negative bin). After sorting more objects in this way, the Find Rules button can again be used to improve classification by learning rules based on the expanded training set.

The accuracy rate can be monitored informally by examining the objects retrieved by CellProfiler Analyst when requested and judging how many are misclassified. For a more formal accuracy assessment, click the “X-Validate” button: it partitions your training set into a specified number of groups, trains on all but one of them, and then using the rules from that training applies them to the left-out set so that a truth table can be built once all groups have been left-out and scored. Note, however, that the accuracy as judged by this method is not the true accuracy of your classifier for the cell population overall. It is merely the accuracy of the classifier on the training set, which likely contains a much more difficult-to-score subset of the overall cell population. It is therefore more like a ‘worst-case scenario’ accuracy rate.

This is a good stage for another type of further iterative training: images containing a high number of positive objects should be examined (by using the “Score Specific Image” button) to see whether the objects score as positive are not false positives. Similarly, images of mostly negative cells can be examined. In either case, further training may be needed to improve accuracy.

Score images
Once you have established a set of rules that achieves an acceptable accuracy rate, you can apply these rules to automatically classify all objects in the experiment by clicking the “Score all images” button. Positive objects can be counted per-image or per-group (groups are specified in the Properties file). Other fields in the output are the total cell count per group, positive cell count per group, the enrichment score (the negative log_{10} of the ratio of the right and left p-values), the right tail p-value, the left tail p-value, and the fold change (percent positives in the group divided by the percent positives in the population). P-values are computed from a Beta-binomial model fit to the entire dataset.

Save/Load Training Sets
When finished, or to save your work and resume later, use the options for saving training sets in the File menu. You may then use Load Training Sets to automatically load these training sets back into classifier. This will automatically empty all classification bins and load your previously classified training set.

Save/Load/Add Rules
When finished, or to save your work and resume later, use the “Save Rule Set” and “Load Rule Set” items in the Classifier File menu. These respectively save and load rules as they appear in the “Train Classifier” panel from/to a text file. Note that saving training sets is much more important than saving rules – rules can be re-generated from a training set, but not vice versa, so you must save a training set as a record of an experiment. Advanced users can also change the rules by typing into the rules window, or add a custom rule to specify criteria.
that an object must meet to be classified as positive, in addition to any other rules that may be developed in the course of building a training set.

Figure 11: Example Classifier Window, with many positive (left) and negative (right) cells classified by the user, creating a large training set. The user has trained the computer using “Find Rules” and has subsequently requested 24 positive cells, which are displayed in the Unclassified bin.