Image Analysis in High-Content Screening

A. Niederlein[1], F. Meyenhofer[1], D. White[2] and M. Bickle[*,1]

[1]*High Throughput Technology Development Studio, Max Planck Institute for Molecular Cell Biology and Genetics, 108 Pfotenhauerstrasse, D-01307, Dresden, Germany*

[2]*Light Microscopy Facility, Max Planck Institute for Molecular Cell Biology and Genetics, 108 Pfotenhauerstrasse, D-01307, Dresden, Germany*

Abstract: The field of High Content Screening (HCS) has evolved from a technology used exclusively by the pharmaceutical industry for secondary drug screening, to a technology used for primary drug screening and basic research in academia. The size and the complexity of the screens have been steadily increasing. This is reflected in the fact that the major challenges facing the field at the present are data mining and data storage due to the large amount of data generated during HCS. On the one hand, technological progress of fully automated image acquisition platforms, and on the other hand advances in the field of automated image analysis have made this technology more powerful and more accessible to less specialized users. Image analysis solutions for many biological problems exist and more are being developed to increase both the quality and the quantity of data extracted from the images acquired during the screens. We highlight in this review some of the major challenges facing automatic high throughput image analysis and present some of the software solutions available on the market or from academic open source solutions.

Keywords: Image quality, segmentation, object features, software.

INTRODUCTION

High Content Screening (HCS) is a term coined by Cellomics in the mid-90s and refers to the high-throughput phenotypic screening of cells using automated microscopes followed by automated image analysis [1]. The basic experimental design of an HCS screen is the following. Cells are cultivated in 96 or 384 wells plates with optical grade plastic bottoms, although alternative culture conditions exist [2]. The cells are then treated either with an arrayed RNAi library or chemical compounds. Cells are fixed and proteins or organelles are visualized using immunofluorescence or recombinant proteins tagged with fluorescent proteins. The plates are placed in an automatic microscope and each well is imaged at several sites. In this manner several hundreds of cells are imaged per well in order to obtain suitable statistics. Such a screen can easily generate millions of images using several Terabytes of storage space. Each imaged cell is then analyzed automatically by image analysis software to identify objects and to measure the size, texture, shape, intensity of fluorescence, location of objects within the cells, spatial distribution and many other parameters. When using multiple labeling, these measurements can be extended to every labeled structure and relationships between the structures can also be calculated. A very large amount of quantitative, spatial data can be extracted from fluorescently labeled cells and the term "high content" in HCS refers to the high density of information such screens can yield. The metadata of such screens can be as large or larger than the primary data itself, taking up again several Terabytes of storage space. In recent years, efforts have been made to develop high content screens using live cells in order to also collect kinetic data and avoid false interpretation of secondary phenotypes observed in end point assays. This will increase the amount of data generated further.

High content screening has become feasible due to the development of automatic image acquisition platforms and the development of automatic image analysis. In HCS, automated image analysis plays a key enabling role as the analysis and interpretation of millions of pictures would simply not be feasible manually. Furthermore, automated image analysis is unbiased, quantitative and reproducible, which is essential to interpret data in a significant manner. The capacity of image analysis to describe features quantitatively is particularly important and should be stressed. The human eye cannot measure absolute intensity values, only relative values are observed and the difference between intensities has to be rather large. Images recorded by a 12 bit camera contain up to 4096 grey levels and very small changes can be recorded that the human eye cannot distinguish. Therefore a computer is required to extract quantitative intensity data. In order for the image analysis to be quantitative, some rules have to be observed when acquiring images and we will discuss some of these rules in this review.

SPECIAL CONSIDERATIONS FOR IMAGE ANALYSIS IN HCS

Given the large size of high content screens, automated image analysis has some particular characteristics in this context. First, the images that are acquired are often of poorer quality than images acquired in a low throughput research mode due to technical restrictions. Cells are often grown in plates with optical grade plastic bottoms instead of glass bottoms to lower costs, but to the detriment of the optical properties. All automatic microscopes are inverted and most use dry long distance objectives, lowering the

*Address correspondence to this author at the High Throughput Technology Development Studio, Max Planck Institute for Molecular Cell Biology and Genetics, 108 Pfotenhauerstrasse, D01307 Dresden, Germany; Tel: +49 (0)351 210 2595; Fax: +49 (0)351 210 1349; E-mail: bickle@mpi-cbg.de

1386-2073/09 $55.00+.00

resolution and image quality. Due to time constraints, it is impracticable to obtain z stacks to ameliorate resolution and contrast by deconvolution and maximum intensity z projection.

All these factors reduce the quality of images produced in high content screens.

Second, there are also restrictions imposed on the image analysis process itself due to the large dataset to be analyzed. The method of analysis has to be applied to a large collection of images of varying quality. It is unavoidable that the preparation of samples and the performance of the microscope will vary to a certain degree during a screen. This, in turn, leads to varying image quality and the image analysis process has to be robust and flexible enough to deal with this variation. Furthermore, unpredictable phenotypes can be encountered in a screen and the image analysis has to be able to deal successfully with a wide range of images containing very diverse information. As automatic segmentation is very difficult, a degree of incorrect object identification must be taken into account. The performance of the image analysis process has to be tested in a pilot screen with known phenotypes to ensure that the experimental setup works and that all the expected phenotypes can be scored.

A screen can easily generate several Terabytes of data and it is therefore important to keep in mind the computational effort involved in the analysis of the data. The analysis time of each picture has to be recorded and should be optimized. Very complicated algorithms that perform better segmentation than simpler methods might not be practical if the computational cost is too high. The need for optimization will depend on the power of the computer cluster used. It should also be kept in mind that image analysis is only the start of the computational part of a high content screen. The measured parameters then have to be compared to each other, clustered and classified using advanced data mining and classification tools. Therefore, computation time per image is an important factor to consider when setting up an image analysis approach in the context of HCS.

IMAGE ACQUISITION CONSIDERATIONS FOR AUTOMATED IMAGE ANALYSIS

The purpose of image analysis in biology is to detect objects and measure parameters that describe the object. The parameters measured can be intensity, size, shape, spatial distribution and many more. In order to fulfill this function, the images that are to be analyzed must be of the highest possible quality (garbage in – garbage out!). The quality is often improved by a preprocessing step to filter out noise, correct for uneven illumination and chromatic aberrations or to enhance edges. But as general rule, out-of-focus objects, uneven illumination or under or overexposed images cannot be reliably analyzed. Given the importance of the image acquisition process for image analysis, we consider it as an integral part of image analysis and will discuss some crucial points.

There exists no generic best solution for image acquisition; each solution is dependent on the biology studied and the questions asked.

HARDWARE CONSIDERATIONS

The choice of the appropriate hardware is the first consideration. Two basic microscopy systems exist for HCS: confocal and widefield. There are several factors that can influence this decision, one of these is resolution, which is calculated with the formulas in Table **1**.

Table 1. Formulas to Calculate Approximately the Resolution of Microscopes Depending on Type, Numerical Aperture and Light Wavelength. D = Resolution, λem = Fluorescence Emission Wavelength, NA = Numerical Aperture of the Objective

Resolution	*lateral*	*Axial*
widefield	$D=0.61*\lambda em/NA$	$D=2*\lambda em/NA^2$
confocal	$D=0.4*\lambda em/NA$	$D=1.4*\lambda em/NA^2$

Confocal microscopes have higher resolution than widefield microscopes, due to the fact that a smaller volume is excited and little out-of-focus light is collected which also results in better contrast. Due to time constraints in HCS, generally only one z plane is imaged. Due to the very small volume imaged in confocal microscopy, objects might either be only partially imaged or totally missed. Thus the size of the object and the type of information to be collected will determine whether a confocal or a widefield microscope is needed.

The acquisition time for collecting the millions of images generated during a screen is crucial to determine the length of the screen and should be therefore minimized. Acquisition time is also to be kept at a minimum when working with live cells to minimize phototoxicity. When performing kinetic experiments, great care should be given that the rate of acquisition is on a comparable scale to the speed of the biological event in order to ensure adequate temporal resolution. Due to the scanning method of excitation and collection, confocal acquisition is often slow, although speed can be increased by using spinning disk systems. Also, due to the reduction of out-of-focus light through two pinholes, much light is lost reducing the signal to noise ratio. This is compensated by increasing the light power and the acquisition time that, in turn, causes problems with photobleaching and phototoxicity. Very fast scanning methods reduce photobleaching and phototoxicity and the fastest scanning method should be chosen, especially for live cell imaging.

Widefield images have higher signal to noise ratio and fast acquisition time but suffer from out-of-focus light and lower resolution. This can be corrected by deconvolution if the point spread function of the microscope is known or can be approximated, but this is computationally expensive and can also lead to artifacts.

The second consideration is the choice of objective and its lens magnification power, numerical aperture, aberration correction and transmission characteristics. The lens magnification power depends on the spatial dimensions of the object to be imaged and what the sampling for statistical purposes should be. For imaging whole cells, counting nuclei

or looking at expression levels of a fluorescent marker, 4x or 10x objectives might be sufficient and allow sampling of large parts of the population with few sites imaged. With fewer sites imaged, the screen takes less time to perform. For sub-cellular resolution, higher magnification lenses are needed.

To judge whether the magnification of the objective and the resolution of the camera is sufficient to image accurately the object of study, the pixel size has to be calculated with the formula in the Table **2**. To accurately image an object, the pixel size should be according to Nyquist theorem of sampling: required resolution/2.3.

Table 2. Formula to Calculate the Pixel Size of a Microscope. *P*size = Edge Length of One Pixel, *CCD*size = Edge Length of One Sensor Field of the CCD Camera, b = Binning, M = Magnification (Lens and Camera Connector)

Pixel Size	*P*size=*CCD*size∗*b/M*

Another important property of the objective is the numerical aperture (NA), which is generally as high as possible to maximize the amount of light collected, the resolution of the image and the field of view obtained. It is also important that the lens is corrected for chromatic aberration to avoid different colors appearing in different z planes or in different x/y coordinates. Most imaging platforms use dry objectives, but the OPERA from Perkin Elmer (formerly Evotec Technologies) uses water immersion objectives, which also improves resolution.

A third consideration is the auto-focus mode of the platform used. Two types of auto-focus are found in most HCS platforms: image-based auto-focus and hardware auto-focus. In image-based auto-focus, images are acquired and an algorithm estimates the sharpness and contrast of the images to find the right focus. In hardware focus, a diode laser is shone at the bottom of the well and the change of refractive index at the transitions of air to plastic and plastic to medium are detected, revealing the position of the bottom of the well. Hardware focusing is fast, but requires that the offset of the objects to be imaged is known, as the fluorescent structures are normally a few microns above the bottom of the plate. Image based auto-focus is slower, leads to photobleaching and phototoxicity and cannot be used on inducible signals. The advantage of image-based focus is that the object of interest is assured to be in focus if used for the focusing process. Some systems combine both the hardware and the image-based autofocus. In this setup, the approximate position of the cells is found by the hardware autofocus and the image-based autofocus only performs the fine focus, which is fast.

Other hardware considerations for good image acquisition, such as stable illumination source and even illumination are essential and should be a given on screening platforms.

IMAGING CONSIDERATIONS

Beside the hardware considerations; the imaging conditions should be optimized in order to obtain the highest possible quality of images to facilitate the image analysis.

The choice of the fluorescent dyes or the fluorescent proteins is critical. Many dyes are commercially available and excitation maxima can be found from 340 nm to 770 nm and emission maxima from 450 nm to 800 nm. When acquiring images with several colors, the excitation and emission spectra have to be carefully analyzed in order to avoid excitation cross talk and emission bleed through due to overlapping excitation and emission spectra of the different dyes. Some websites offer the possibility to look at the excitation and emission spectra of many dyes and fluorescent proteins to verify the overlap of the spectra (i.e. http://probes.invitrogen.com/resources/spectraviewer/). The excitation filters or laser lines and the emission filters fitted on the imaging platform must also be compatible with the excitation and emission spectra of the dyes.

Auto-fluorescence of biological samples is much reduced above 600 nm, resulting in higher signal to noise ratio. A drawback is, that at longer wavelengths, the resolution is also reduced. For living samples, heating of the sample, resulting in photodamage, can become an issue at wavelengths above 700 nm. Wavelengths in the range of 450 to 550 nm are very well detected by most CCD cameras found on microscopes, but autofluorescence can be an issue for weak signals. At shorter wavelengths, mercury light sources emit very strongly, exciting thereby the fluorophores very strongly, but short wavelengths are very harmful for living biological samples, resulting again in photodamage.

Lastly, the quantum yield of the fluorophore (amount of photons emitted per photons received) and the stability of the signal over time are important characteristics of the dye to be considered.

The brightness of the objects imaged is very important for the image analysis process. There should be no saturated pixels and the background while being near zero should be above it. In this manner, one ensures that details in bright and faint regions can be recorded and analyzed. The brightness is controlled in confocal microscopy by the laser power, the aperture of the pinholes and the exposure time. In widefield microscopy acquisition time can be controlled and sometime also illumination power. A digital microscopic image is characterized by a finite number of picture elements (pixels) representing the fluorescence on an area of a defined length, width and depth (axial / lateral resolution). Each pixel records a grey level value corresponding to the brightness of the corresponding area in the imaged object. The dynamic range of the gray level values depends on the camera used. Most scientific grade CCD cameras are 12 bit and therefore can record grey level values from 0 to 4095. It is important when acquiring images that most of the dynamic range of the camera is used. This does not necessarily result in an image satisfying to the eye, but it is important to realize that images taken for quantitative analysis do not necessarily satisfy our esthetic senses.

The demands for correct exposure can be in conflict with the throughput requirements of HCS. Speed is an important factor when considering screening procedure and setup. Images have to be acquired with the minimal exposure time that can still satisfy the image analysis requirement in order to reduce the overall time of the screen. Consider a genome-wide RNAi screen where 25000 genes are to be analyzed using three double stranded RNA oligonucleotides per gene.

If two color images at 15 sites per well are recorded, 2.25 million images will be collected. An exposition time difference of 100 ms will add nearly 3 days to the screen. With more complex screens, with more colors, larger libraries or in kinetic mode, the time scale becomes correspondingly larger. Furthermore, when taking images of live cells, the rate frame required to follow a biological event with sufficient accuracy might impose constraints on the exposure time. In summary, a compromise between minimal exposure time and sufficient signal strength for image analysis has to be found.

TYPICAL STEPS OF AUTOMATED IMAGE ANALYSIS

A problem specifically encountered in HCS, is images that either contain no, few or out of focus objects. This is due to uneven cell seeding and spreading, toxic experimental conditions and occasional errors in the autofocus method. Some automated microscopes have on the fly quality control to ensure that images are only acquired when objects are found. This reduces the amount of empty images. For blurred images, the image analysis process has to incorporate a quality control step to eliminate these images from the subsequent analysis in order to avoid false interpretations due to erroneous segmentation.

After the quality control step, the process of image analysis follows generally the same procedure whether working with large or smaller datasets. In the first instance, the images that are acquired will be corrected if required. As it is difficult to obtain even illumination in microscopy, background correction might have to be applied, especially when analyzing fluorescence intensities. Post-acquisition processing might also be required to sharpen some edges, increase contrast or eliminate noise. If possible, it is best to work with raw data, as there is always some concern that the correction process might alter the data. The next step is then to find objects or regions of interest within the images. This process is called segmentation and is the most difficult step of image analysis. The human brain is very good at finding patterns even within very noisy images, but computers do not have this faculty. Once the objects have been successfully identified, features are extracted that describe the objects themselves, such as shape and texture, and relationship amongst objects can also be calculated, such as distance and distribution. The image analysis process is then finished. The classification and interpretation of measurements are not part of the image analysis itself and will not be discussed here.

Correction of Images

In order to make the distinction between objects and background more evident and removing parasite signals (noise), most image processing routines begin with a step called background correction or preprocessing. This first step can be broken down into two different types of actions: subtraction and normalization of the background and noise filtering. The background subtraction is prone to alter the signal information and therefore is only done if necessary.

Background subtraction aims at flattening uneven illumination and at eliminating the offset and so pulls the background closer to the zero value of the gray-scale. There are many methods to perform this operation (for review see reference [3]). The simplest way is to subtract the mean value of the background from the whole image. This is achieved by identifying an area of the image as background, calculating its mean value and subtracting that value from the value of each pixel in the entire image. This approach is very efficient in images with relatively few objects as the mean of the background values is similar to the mean of the values of the whole picture; the offset is drastically reduced. Another approach consists of fitting a polynomial to some background sample pixels in the least square sense, then subtracting the obtained polynomial approximation of the background from each pixel of the image. A third approach, based on the shape of the objects is the subtraction of the morphological opening from the image (morphological top hat [4]). The opening of the image filters out the objects and leaves the smoothed background image. It is very effective if the objects are small and their density not too high.

Common noise filters are the mean, median and Gaussian filters for statistical noise (Gaussian, uniform or salt and pepper noise). In case of periodical noise, working in the frequency space and applying band pass and band reject filters can be very effective for object enhancement.

To obtain the best result in the preprocessing step it is essential that the methods hypothesis fits the given problem and that its parameters are carefully adapted. Thus, applying a band pass filter to random noise can introduce artifacts, and when applying a Gaussian filter the size of the kernel has to be carefully chosen so as not to blur edges too much.

Finally, objects that are on the border of the image and are incomplete have to be filtered out of the analysis.

Segmentation

Once the image fulfills appropriate quality criteria, objects are recognized using various segmentation algorithms. This part of the image analysis process is the most difficult and also arguably the most crucial step for the success of the analysis. The easiest method for segmentation is manual where the user draws the outline of the object. This is obviously not applicable to HCS. Many automated segmentation techniques exist and more are being developed every year. Segmentation can be helped by setting up the experiments in such a way, that the objects are not too crowded for object identification, and the illumination is flat giving good contrast.

The simplest method of segmentation is called thresholding and uses the histogram of grey level values of the pixels (intensity histogram). Many different thresholding algorithms have been developed to set the threshold value (reviewed in reference [5]). These methods can be applied over the entire image if the illumination is even, the full dynamic range has been used and the staining of the objects is reproducible, resulting in good contrast and high signal to noise ratio. If these conditions are met, two peaks should be evident in the intensity histogram and the threshold is set at the bottom of the valley between the two peaks. Adjacent pixels with values above the threshold are considered as belonging to an object and below the threshold as belonging to the background A very popular algorithm is Otsu's method, which determines the threshold by maximizing the variance between the background and the objects (inter-class

variance) and minimizes the variance within objects (inner-class variance) [6]. This is a very simple method but will often fail to segment clustered objects or images with many intensity values. If the images have uneven illumination local thresholds based on spatial and gray-scale intensity have to be determined [7].

Another widely used class of segmentation methods is based on edge detection. Edges of an object are characterized by local changes in pixel intensity with a sharp gradient. Edge detection algorithms look for such sharp gradients using for instance Sobel or Canny edge detection.

One popular algorithm using gradient information is the watershed transformation [8]. The idea of this morphological segmentation method is to consider the gradient distribution as a topographical landscape with high image gradient areas forming peaks and low gradient areas forming valleys and plains. Holes are drilled in the minima and the imaginary landscape is slowly sunk into water. As the water starts rising from the local minima, the valleys are flooded and water from adjacent valleys meet. The watershed segmentation lines are drawn along the line where water meets. Depending on the setting of the parameters, objects can be under- or over-segmented and the method has seen a lot of modification and adaptation to specific applications. For instance, merging of oversegmented regions according to shape and intensity similarities can be applied. Those types of segmentation are fast and thus well suited to HCS.

More sophisticated segmentation approaches based on gradient images like active contour or the LiveWire algorithm use edge tracing [9]. The idea of region based-segmentation is to set some seed objects or contours. The object are then optimized or grown until the optimum of some object criteria is reached. The growth is restricted by rules concerning, for instance, the homogeneity of the object (regions of similar intensity). Such an approach is described by Yan and colleagues for correctly segmenting cell bodies in HCS context [10]. First, nuclei segmentation is performed by simple thresholding. Second, a distance transform map can be applied to which the watershed method is then applied. The two results can be merged to obtain good segmentation of the nuclei. The locations of the nuclei are then used as seeds to look for the cell contours using contrast, gradient and intensity information in a repulsion and competition model. Due to the iterative nature of the approach, it is computationally expensive.

A further method of segmentation is based on modeling of the objects to be detected. Either a template image is used or the shape of the object is described by a mathematical model with several variables. Intensity values are looked for in the image that fit the model. This type of approach can also be computationally expensive (reviewed in reference [11]).

As mentioned above, segmentation is the most difficult and most crucial part of image analysis and researchers try applying different techniques to their specific problems for comparison [12, 13]. Combinations of all the above-mentioned methods can also be applied to improve segmentation of the objects. In many instances, it is found that the basic segmentation algorithms do not give satisfying results and that further refinement has to be brought to the initial segmentation result. It is also possible to apply morphometric parameters and feed back control if the shape of the object can be modeled [14].

A generic solution does not exist for segmenting images and much development is still required in order to refine the existing solutions, adapt and establish basic and flexible tools for fluorescent high content screening. The researcher normally has to decide what level of erroneous segmentation resulting in objects being lost, fused or split is acceptable.

Feature Extraction

Once objects have been successfully identified, quantitative features are then extracted. A surprising amount of features can be extracted from images some of which are evident to a biologist some of which are not [15, 16]. For instance, measuring the intensity of a DNA stain to assess DNA content is obvious, whereas measuring texture features of a nucleus is not. Nevertheless, it has been found that the more features are extracted, the more information can be gained to help to classify phenotypes.

Here follows a list of some of the features that can be quantified following object identification:

- Intensity features: to quantify the amount of labeled protein
  - Total intensity
  - Mean intensity
  - Median intensity
  - Distribution of intensity
- Shape features: to describe the morphology of object
  - Size of the perimeter
  - Length of major axis
  - Length of minor axis
  - Roughness of the perimeter
  - Amount of convex or concave structures
  - Length of the convex and concave structures
- Surface features: to describe area of fluorescence
  - Total area
  - Mean area
  - Median area
  - Center of mass
  - Centroid of mass
  - Texture
- Distribution features: to describe area and distribution of fluorescence
  - Mean peer to peer distribution
  - Median peer to peer distribution
  - Mean distance to other objects
  - Median distance to other objects

- Whole image features: to describe the number and distribution of objects
    - Total number of objects
    - Number of sub objects within objects

This list is by no means exhaustive. These features are extracted for each fluorescent channel, resulting in a very large amount of data for each experimental condition. Once the feature extraction is completed, the image analysis process *per se* is finished and a large amount of metadata has been created. The next step is to classify the feature sets of the objects of interest and relate them to the experimental treatment. This is beyond the scope of this review and concerns data mining approaches for classification.

IMAGE ANALYSIS SOFTWARE

All HCS imaging platforms come equipped with software for image acquisition control, image analysis, statistical tools, data visualization tools and a database for storing and retrieving data. The image analysis solutions found on HCS imaging platforms were developed primarily for the pharmaceutical industry. They offer image analysis solutions for some of the typical screening requirements in the drug discovery industry, such as protein translocation, micronuclei detection, neurite outgrowth, tube formation, cell count and target activation. These algorithms are designed to be tunable to adapt to screening in different cell lines with different markers. They are meant to be usable with minimal training on the part of the user and do not require the user to develop his own tools or to be able to program. One drawback of such ready-to-go systems is their lack of flexibility concerning the type of biological problems that can be tackled. Additionally, as the source code is protected, the user does not know the methodology of the different steps, cannot modify it or fully understand its functionality. Also, proprietary software licences are expensive and may not be affordable for all users.

More flexible platforms exist and can be divided in two categories: open source and proprietary. Proprietary softwares are for instance MatLab and MetaMorph, whereas open source softwares are ImageJ and Cellprofiler (Carpenter 2006).

In the following sections, we will discuss some features of the most commonly available softwares on the market. More software packages exist, but will not be discussed here.

Image J

Image J is an open source software that has been developed using the Java programming language. It is freely downloadable at http://rsb.info.nih.gov/ij/. It has been developed by the user community and there are over 300 plugins available that fulfill various functionalities. It can run on UNIX, Linux, Mac OS X and Windows platforms. Plugins exist for import of most file formats making Image J compatible with probably all imaging platforms. It is also possible to add or modify plugins if cognizant of the Java programming language. Due to the large documentation available on the internet and graphical user interfaces (GUIs), non-computer scientist can easily learn how to use it. As a large user community exists that can be contacted *via* a mailing list, help can be found when a user encounters a problem. Macros can easily be created, allowing the automation of the image analysis process, which is essential for HCS. Some built-in applications such as 'MRI Cell Image Analyzer' are also available.

Matlab

Matlab is a commercially available, high-level computing language for algorithm development, data visualization, data analysis and numeric computation. It can be used for many applications, amongst which image analysis is but one. Although Matlab is proprietary, applications using Matlab languages are freely available (Cellprofiler [17], CellC [18]). It can interface with other software, deal with most of the common format types and versions are available to run on Mac OS X, Windows or Linux.

It is not an easy-to-use application and requires a computer scientist to operate it. Toolboxes are available to help in the development of novel algorithms. As a developer tool, it is highly flexible but also requires time to optimize the image analysis process. Programming is fast in Matlab even if the language itself is not as fast as C++. Matlab can also be designed for distributed computing to master large amounts of data in reasonable time.

Cellprofiler

Cellprofiler is an open source application based on Matlab that has been specifically developed for HCS [17]. It is freely downloadable at http://www.cellprofiler.org in versions compatible with Mac OS X, Windows and Unix. Matlab is not required to be installed to run the application, but the source code for Cellprofiler is also downloadable on the website for Matlab users. It is possible to write more code for the application in order to expand the capacities of the program and it also possible to modify the code to adapt to specific problems.

Cellprofiler has been specifically designed to bridge the gap between developer tools such as Matlab and the proprietary software for HCS. It offers approximately 50 modules for typical image analysis steps with user-friendly GUIs. Many image file formats can be read and, if necessary, code can be written to accept further formats. An image analysis project is constructed as a pipeline, where several modules, each carrying out an image analysis step, are set up sequentially in order to process images. The user can modify parameters to adapt the algorithms to the specific task at hand. If no module can do the required job, the user can develop his own algorithm, provided he is proficient in Matlab. Cellprofiler is relatively new, but a user community is growing so that new modules should be appearing at an increasing rate.

As Cellprofiler was designed for HCS, distributed computing is feasible, so that clusters can be used.

Definiens Cellenger

Definiens Cellenger is a commercially available image analysis software designed for HCS that is independent of an imaging platform [19, 20]. It runs under Windows exclusively and supports most HCS image formats. The focus of the software is complex image analysis. The software was created to offer on the one hand easy-to-use

tools for non-specialists, while on the other hand allowing more specialized users to develop their own tools.

The Definiens Architect, allows the biologist to use some robust modules. Only a few parameters can be adjusted and a comfortable user interface is provided to check the correctness of the parameters. Ready-made solutions exist for most common applications such as nuclei detection, translocation assay, tube formation.

Developer is the more advanced image analysis package for computer scientists. The image analysis process is built from so-called rulesets, which are sequential arrangements of single processes. The basic principle is an iterative process of segmentation, classification and merging. It is an object-based approach, which means that every part of the image – created by segmentation and merging steps – will be handled as an object. The software exploits several segmentation techniques and extracts a very large amount of data out of the raw measurements. The cleverness of the software is its faculty to break down the image into several objects that then can be merged for accurate segmentation. In the most extreme case, every pixel can be segmented as an object; neighboring pixels can then be merged according to intensity and cellular objects thereby reconstructed. Additionally, the software provides hierarchical levels of refinement. This enables the detection and feature extraction of objects like the cell and its compartments at the same time. The classification of objects is based on object features, which range from simple features like intensity features, to shape features and very useful relational features (Relations to neighbor, child or parent objects). A complete ruleset then can be organized like a module for the Architect, so that biologist could also make use of these solutions in a comfortable way.

The latest version of the software allows now also 3D and 4D image analysis with object tracking capacity.

One drawback of the software is that there is only one segmentation layer and it is not possible to draw object masks for each channel separately. The software also allows visualization of the data cell by cell, well by well and plate by plate with statistical analysis (means, standard deviation, z-factor).

The software is able to run the analysis on multiple remote machines. These analysis engines as well as the user interface are licence- restricted. Batch processing is simple and comfortable to set up and results can also be displayed by plate heatmaps.

Acapella

This software is supplied with the spinning disc confocal microscope OPERA of Perkin Elmer (formerly Evotec Technologies) [21]. Acapella runs on Windows and can make use of multiple analysis engines. These engines are licence-restricted. The analysis can run on-the-fly during acquisition, as well as separately. Images acquired by the Opera can run very easily as batch jobs producing output parameters that can be visualized as heatmaps. This kind of visualization is useful for fast, automated quality control during a screen. Acapella provides the read-in of standard image formats, so that images taken by other acquisition instruments can also be analyzed.

Image analysis runs on user-written scripts. These scripts consist of modules provided by different module libraries. The Acapella user interface allows the creation of scripts by drag and drop (module-tree within a so-called "block-editor") as well as by editing within the text editor. The latest release of Acapella has improved the user interface and the software is now more user friendly.

The modules range from very sophisticated modules specific to HCS, to basic image analysis modules. There are four libraries providing modules for very fast and flexible analysis of cell, nuclei and spot detection. Selection modules help to choose the appropriate algorithm. It is important to test different conditions (positive/negative controls, untreated) to guarantee robustness. With such modules, an image analysis task – depending on complexity – can be done with a very small script and programming skills are not obligatory. Programming on a high level using C++ is also possible if the need arises for a specific task that is not covered by the modules or where the modules perform poorly. Acapella provides an interface to extend the functionality. McMaster University hosts Acapella scripts developed in house on their website and these scripts are downloadable (http://www.macbiophotonics.ca/downloadsa.htm). Objects are managed by object lists that allow comfortable definition of object features.

BioApplication

Is a commercial software from Cellomics (now part of Thermo Fisher Scientific) that is installed on their imaging platform, the ArrayScan VTI. Cellomics has dominated the HCS market in the past 10 years and many successful screens have been realized with their technology [22-25]. Every BioApplication is made for a specific kind of assay and parameters are adjusted to obtain correct segmentation according to the experimental setup. As the functionality of the BioApplications were designed for specific assays, adaptation to other biological problems is more difficult, but quite feasible for experienced users. It is designed as a turnkey solution for users with little experience in image analysis. The parameters within the BioApplications can be adjusted by clicking on objects and reading their values. These values help set the upper and lower limits of the parameters used in the BioApplication. A drawback of these simple solutions is that the user does not know what the algorithm calculates. It is therefore important to test the algorithm under many experimental conditions to ascertain that it will perform correctly under the many conditions encountered in a screen. For instance, depending on the type of background correction applied, the surface of nuclei might appear to be dependent on the density of the cells. This might actually be an artifact due to the fact that the background increases with cell density and the background correction alters the segmentation of the nuclei.

The algorithm of BioApplications like "Cell Cycle", "Cytoplasm to Nucleus Translocation" or "Neurite Outgrowth" provides quite simple and fast image analysis. The focus is less on accuracy than on speed and statistic significant readout from a large amount of data analyzed. It runs on Windows and only analyzes images acquired by the ArrayScan. Recently, environmental control has been added

to the ArrayScan and the software has been upgraded to also analyze kinetic data.

AttoVision

AttoVision is the proprietary software with which the BD Pathway 855 Bioimager (BD Biosciences) imaging platform is equipped. This platform has a promising potential due to the flexibility of the illumination settings, the optional confocality, the injection capability, the incubation chamber and the high quality optics [26, 27]. A new version of the software was released in 2008 allowing more control of the image analysis process. Furthermore, the new version is capable of more complex segmentation tasks such as multiple bands and rings and allowing more complex tasks. The GUI integrates the image capture, analysis and interpretation. On-the-fly analysis and classification using Boolean logic is also provided. Even though the BD Pathway is a microscope with environmental control, the software is unable to treat kinetic data correctly. The segmentation is only performed on the first image of a stack and the mask is used in all subsequent frames. Thus, if an object moves, the mask will be wrong and the measurement meaningless.

IN Cell Investigator

IN Cell Investigator is the proprietary software of GE Healthcare and is installed on their IN Cell Analyzer 1000 and 3000 microscopes. It also has easy-to-use tools and developer tools for more sophisticated application requiring more image analysis knowledge. It offers 10 different modules that can be applied to many biological questions [28-30]. It also has a Decision Tree filter for classifying objects according to a specified measurement. There are multiple levels of decision and classification for sorting subpopulations. The developer toolbox is designed for unskilled users to customize image analysis routines. For visualizing the results of the image processing, a Spotfire DecisionSite Basic software is included. The software can also accept other file formats than its own by purchasing the IN Cell Translator Software.

MetaXpress

MetaXpress is the proprietary software of Molecular Devices and equips its three imaging platforms: ImageXpress Micro, ImageXpress Ultra and ImageXpress 5000A. So-called Application Modules can be purchased for specific tasks, such as cell cycle analysis, tube formation, nuclei counting etc [31]. There are easy-to-use and do not require any programming skills. No real developer function is available for specialized users.

Kalaimoscope

This software is newly commercially available and was developed in the Max Planck Institute of Molecular Cell Biology and Genetics (MPI-CBG) of Dresden, Germany, by Prof. Yannis Kalaidzidis. It was originally developed to detect endosomes and to track them over time during their maturation [32]. The originality of the software lies in the fact that detection of objects is based on mathematical modeling. Each intensity peak gives a seed to fit the parameters of the model resulting in a round or elliptic representation. Multiple simple objects can afterwards be combined to more complex structures allowing the detection of other intracellular objects such as nuclei, mitochondria and single fluorescent molecules. With this approach small structures like endosomes can be detected with subpixel precision, allowing detection and localization of objects beyond the resolution capacity of the microscope. Even phenotypic changes in size can be measured, even though the signal of the objects may only be present on a few pixels. The software is also capable of segmenting objects that have a very low signal to noise ratio, either due to faint labeling or high background. The parameters that have to be adjusted by the user are few and pertain essentially to the object size and intensity. The tuning of these parameters can be done on one single representative image.

Once an object has been detected, the software assigns an *x-y* position. When doing time series the software links the positional information of the objects between the timeframes allowing tracking of each object over time.

Furthermore, the software allows sophisticated preprocessing and correction steps and offers advance statistical tools. All in all, around sixty parameters can be extracted for each object. The disadvantage of the software is the computational cost. To handle the time-consuming object detection method and their feature calculations, the application is able to run on a Linux PC cluster and it is recommended to have many CPUs available. The MPI-CBG takes advantage of the super computer of the Technische Universität of Dresden to run image analysis on up to 2400 CPU.

CONCLUSION

Image analysis is still a bottleneck in the field of HCS. This is due to several problems. First, image quality in HCS is often poor compared to low throughput applications, due to specific screening constraints. Second, due to the variability of biological structures regarding size, shape and intensity in screens, efficient automatic segmentation procedures are difficult to devise. A certain amount of false object identification has to be taken into account, forcing the researcher to image more cells in order to obtain statistically significant data. Third, due to time constraints, computationally expensive image analysis solutions are only available to researchers with access to very large computer clusters.

Many software solutions are now available for image analysis in HCS context, ranging from simple turnkey solutions to more complex expert user solutions. As the field progresses and more screens are successfully completed, more and more image analysis solutions will appear, both in the open source environment and commercial solutions. This should allow users to find solutions that will require only a little tuning to be adapted to their biological problem. With the development of faster chips and the creation of local clusters using internal networks in universities and companies, more sophisticated and computationally expensive image analysis solutions should become practicable.

One field of growth in HCS image analysis deals with kinetic studies. Software packages that can cope with kinetic data are rare and few screens have been published. As live cells screens are developed, algorithms to analyze the kinetic data will emerge in the future.

ACKNOWLEDGEMENTS

We are deeply indebted to Dr. Eberhard Krauß, Dr. Eugenio Fava, Prof. Yannis Kalaidzidis and Dr. Anne Eugster for critical reading of the manuscript. This work was supported by the grant 0313831A from the Bundesministerium für Bildung und Forschung (BMBF), the FP6 Endotrack program and the Max Planck Society.

REFERENCES

[1] Giuliano, K.A.; DeBiasio, R.L.; Dunlay, R.T.; Gough, A.; Volosky, J.M.; Zock, J.; Pavlakis, G.N.; Taylor, D.L. High-content screening: a new approach to easing key bottlenecks in the drug discovery process. *J Biomol Screen.* **1997**, *2*, 249.

[2] Erfle, H.; Neumann, B.; Liebel, U.; Rogers, P.; Held, M.; Walter, T.; Ellenberg, J.; Pepperkok, R. Reverse transfection on cell arrays for high content screening microscopy. *Nat Protocols,* **2007**, 2, 392.

[3] Russ, J.C. *The Image Processing Handbook*. 5th ed.; CRC Press: **2007**; p 818.

[4] Meyer F, Beucher S. Morphological segmentation. *J Visual Commun Image Represent.* **1990**, *1*, 21.

[5] Sahoo, P.K.; Soltani, S.; Wong, A.K.C. A survey of thresholding techniques. *Computer Vision Graphics Image Process.* **1988**, *41*, 233.

[6] Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans Sys Man Cyber.* **1979**, 9, 62-66.

[7] Burghardt, A.; Kazakia, G.; Majumdar, S. A local adaptive threshold strategy for high resolution peripheral quantitative computed tomography of trabecular bone. *Ann. Biomed. Eng.,* **2007**, 3*5*, 1678-1686.

[8] Beucher S. In *The watershed transformation applied to image segmentation*, 10th Pfefferkorn Conference on Signal and Image Processing in Microscopy and Microanalysis, Cambridge, UK, 1992, 1991; Cambridge, UK, **1991**, 299-314.

[9] Barrett, W.A.; Mortensen, E.N. Interactive live-wire boundary extraction. *Med. Image Anal.,* 1997, 1, 331-341.

[10] Yan, P.; Zhou, X.; Shah, M.; Wong, S.T.C. Automatic segmentation of high throughput RNAi fluorescent cellular images. *IEEE* **2007**, 1-8.

[11] Kalaidzidis, Y. Intracellular objects tracking. *Eur J Cell Biol.* **2007**, *86*, 569.

[12] Drever, L.A.; Roa, W.; McEwan, A.; Robinson, D. Comparison of three image segmentation techniques for PET target volume delineation. *J. Appl. Clin. Med. Phys.,* **2007**, *8*, 93-109.

[13] Geets, X.; Lee, J.; Bol, A.; Lonneux, M.; Grégoire, V. A gradient-based method for segmenting FDG-PET images: methodology and validation. *Eur. J. Nucl. Med. Mol. Imaging*, **2007**, *34*, 1427-1438.

[14] Li, F.; Zhou, X.; Ma, J.; Wong, S.T.C. An automated feedback system with the hybrid model of scoring and classification for solving over-segmentation problems in RNAi high content screening. *J. Microscopy,* **2007**, *226*, 121-132.

[15] Paran, Y.; Ilan, M.; Kashman, Y.; Goldstein, S.; Liron, Y.; Geiger, B.; Kam, Z. High-throughput screening of cellular features using high-resolution light-microscopy; Application for profiling drug effects on cell adhesion. *J. Struct. Biol.,* **2007**, *158*, 233-243.

[16] Bakal, C.; Aach, J.; Church, G.; Perrimon, N. Quantitative morphological signatures define local dignaling networks regulating cell morphology. *Science,* **2007**, *316* (5832), 1753-1756.

[17] Carpenter, A.; Jones, T.; Lamprecht, M.; Clarke, C.; Kang, I.; Friman, O.; Guertin, D.; Chang, J.; Lindquist, R.; Moffat, J.; Golland, P.; Sabatini, D. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome. Biol.,* **2006**, *7*, R100.

[18] Selinummi, J.; Seppälä, J.; Yli-Harja, O.; Puhakka, J.A. Software for quantification of labeled bacteria from digital microscope images by automated image analysis. *Biotechniques,* 2**005**, *39*, 859-863.

[19] Biberthaler, P.; Athelogou, M.; Langer, S.; Luchting, B.; Leiderer, R.; Messmer, K. Evaluation of murine liver transmission electron micrographs by an innovative object-based quantitative image analysis system (Cellenger). *Eur. J. Med. Res.,* **2003**, *8*, 275-282.

[20] Baatz, M.; Arini, N.; Schäpe, A.; Binnig, G.; Linssen, B. Object-oriented image analysis for high content screening: Detailed quantification of cells and sub cellular structures with the Cellenger software. *Cytometry Part A* **2006**, *69*A, 652-658.

[21] Garippa RJ, Hoffman AF, Gradl G, Kirsch A. High-throughput confocal microscopy for beta-arrestin-green fluorescent protein translocation G protein-coupled receptor assays using the Evotec Opera. *Methods Enzymol.,* **2006**, *414*, 99-120.

[22] Gasparri, F.; Mariani, M.; Sola, F.; Galvani, A. Quantification of the proliferation index of human dermal fibroblast cultures with the ArrayScanTM high-content screening reader. *J Biomol Screen.* **2004**, 9, 232-243.

[23] Vogt, A.; Cooley, K.A.; Brisson, M.; Tarpley, M.G.; Wipf, P.; Lazo, J.S. Cell-active dual specificity phosphatase inhibitors identified by high-content screening. *Chem Biol.,* **2003**, *10*, 733-742.

[24] Liu, D.; McIlvain, H.B.; Fennell, M.; Dunlop, J.; Wood, A.; Zaleska, M.M.; Graziani, E.I.; Pong, K. Screening of immunophilin ligands by quantitative analysis of neurofilament expression and neurite outgrowth in cultured neurons and cells. *J. Neurosci. Methods,* **2007**, *163*, 310-320.

[25] Trask, J.O.J.; Baker, A.; Williams, R.G.; Nickischer, D.; Kandasamy, R.; Laethem, C.; Johnston, P.A.; James, I. Assay development and case history of a 32K-biased library high-content MK2-EGFP translocation screen to identify p38 mitogen-activated protein kinase inhibitors on the ArrayScan 3.1 imaging platform. *Methods Enzymol.,* **2006**, *414*, 419-439.

[26] Chan, G.K.Y.; Richards, G.R.; Peters, M.; Simpson, P.B. High content kinetic assays of neuronal signaling implemented on BDTM pathway HT. *Assay Drug Develop Technol.,* **2005**, *3*, 623-636.

[27] Zanella, F.; Rosado, A.; Blanco, F.; Henderson, B.R.; Carnero, A.; Link, W. An HTS approach to screen for antagonists of the nuclear export machinery using high content cell-based assays. *Assay Drug Devel. Technol.,* **2007**, *5*, 333.

[28] Ramm, P.; Alexandrov, Y.; Cholewinski, A.; Cybuch, Y.; Nadon, R.; Soltys, B.J. Automated screening of neurite outgrowth. *J. Biomol. Screen,* **2003**, *8*, 7-18.

[29] Lundholt, B.K.; Linde, V.; Loechel, F.; Pedersen, H.C.; Moller, S.; Praestegaard, M.; Mikkelsen, I.; Scudder, K.; Bjorn, S.P.; Heide, M.; Arkhammar, P.O.; Terry, R.; Nielsen, S.J. Identification of Akt pathway inhibitors using redistribution screening on the FLIPR and the IN Cell 3000 Analyzer. *J Biomol Screen.* **2005**, *10*, 20-29.

[30] Granas, C.; Lundholt, B. K.; Heydorn, A.; Linde, V.; Pedersen, H.-C.; Krog-Jensen, C.; Rosenkilde, M. M.; Pagliaro, L. High content screening for G protein-coupled receptors using cell-based protein translocation assays. *Comb Chem High Throughput Screen.* **2005**, *8*, 301-309.

[31] Galvez, T.; Teruel, M.; Heo, W.; Jones, J.; Kim, M.; Liou, J.; Myers, J.; Meyer, T. siRNA screen of the human signaling proteome identifies the PtdIns(3,4,5)P3-mTOR signaling pathway as a primary regulator of transferrin uptake. *Genome Biol.* **2007**, 8, R142.

[32] Rink, J.; Ghigo, E.; Kalaidzidis, Y.; Zerial, M. Rab conversion as a mechanism of progression from early to late endosomes. *Cell,* **2005**, *122*, 735-749.

Received: ?????????? ??, ???? **Revised: ????????? ??, 2008** **Accepted: ???????? ??, ????**