DRUG DISCOVERY TODAY BIOSILICO

Analyzing and mining image databases

Thomas Berlage

Image mining is the application of computer-based techniques that extract and exploit information from large image sets to support human users in generating knowledge from these sources. This review focuses on biomedical applications, in particular automated imaging at the cellular level. An image database is an interactive software application that combines data management, image analysis and visual data mining. The main characteristic of such a system is a layer that represents objects within an image, and that represents a large spectrum of quantitative and semantic object features. The image analysis needs to be adapted to each particular experiment, so 'end-user programming' will be desirable to make the technology more widely applicable.

Image mining is the application of computer-based techniques that extract and exploit information from large image sets to support human users in generating knowledge from these sources (for a general overview, see [1]). This review focuses on biomedical applications, in particular automated imaging at the cellular level. Experiments at this level provide more biological context than molecular approaches, and are more accessible than imaging at the level of complete organisms. Different molecular markers can be observed in thousands of cells under a variety of conditions, and their spatial and temporal patterns can be analyzed per tissue region, per cell or per cellular compartment. Multiplexing techniques are often employed to equalize experimental conditions, for example, by mixing different cell lines [2,3], by using tissue microarrays with material from multiple samples [4,5] or by mixing multiple samples mounted on special carriers with individual optically readable codes [6]. Each image may therefore have a complex substructure.

Automated image acquisition requires special equipment, such as automated microscopes for microtiter plates [7–10] (for a list of automated imaging companies, see [11]). Also needed is an 'image database', an interactive software application that combines data/image management, data/image analysis and data/image visualization [2,5,12–14]. These three aspects are part of every image mining approach. Most commercial equipment comes with its own image database software, although independent software applications are becoming more and more available.

Image analysis plays a central role regarding software, as it derives detailed qualitative and quantitative information from the images. This is often referred to as 'high-content analysis'. A major challenge of image mining is the programming required. Depending on the particular experiment, the user has to define the objects and features to be extracted before the analysis procedure is automatically applied to all images. There is a significant semantic gap between automatic image analysis and a scientist's interpretation [15], although there is also potential for quantitative measurements beyond the capabilities of human perception.

Thomas Berlage
Fraunhofer Institute for Applied Information Technology (FIT),
Schloss Birlinghoven,
53754 Sankt Augustin,
Germany
e-mail: thomas.berlage@fit.fraunhofer.de

1359-6446/05/$ – see front matter ©2005 Elsevier Ltd. All rights reserved. PII: S1359-6446(05)03462-8

A further challenge lies in goal-oriented visualization for interpretation and knowledge formation. Only in a few cases (such as screening with standardized quantitative assays) is the derived information sufficient without further analysis. In the majority of applications, the user needs interactive visualization to assess the quality of the experiment (and the programmed analysis) and to put the wealth of observations in an interpretatory context. Visual data mining lets the user add information through human interpretation of images, but also presents high-level patterns visually. The importance of this stage for efficient analysis is often underestimated.

Mining experimental images requires associated metadata. Without reference to the sample and its properties, and to the experimental conditions and procedures used, image analysis is meaningless. Therefore, data management and data modeling (i.e. 'programming' data structures and representations) are an integral part of every image mining approach [1,4]. Data management poses its own challenge in integrating different data sources. Coping with the sheer amount of image data is no longer a serious issue in the days of the multimedia PC. The bottleneck today is putting together images, image-derived information and metadata from the experiment, and information from public databases to enable the analysis of global patterns.

The next sections will deal with image analysis, visualization and data management individually, followed by a discussion of application areas and future developments.

Image analysis

Defining image analysis procedures is the most time-consuming part of image mining, as the kinds of images as well as the features requested vary among experiments. Biomedical images are based on two-dimensional sections, and mostly do not have to cope with distance and perspective like other multimedia applications (photography, film). Three-dimensional imaging [16] and temporal sequences [17,18] play an increasing role. Furthermore, many images are multimodal, that is to say, there are multiple measurements/channels per object imaged [19].

Hsu *et al.* [1] distinguish the following levels in image mining: pixel level (corresponding to raw and filtered images), object level (corresponding to regions or objects recognized in images), semantic concept level (corresponding to categories or other semantic annotations of the objects), and pattern and knowledge level (corresponding to higher level concepts, structures and relations that link different facts). As mentioned above, the object level is a key information provider in biomedical images. The object level is established through segmentation. A large variety of properties or features can then be calculated for objects. In many applications, automatic analysis stops here, and any further information is established interactively using visualization and statistical tools. However, it is now increasingly possible to address the semantic concept level by performing object classification automatically for the segmented regions, which significantly simplifies dealing with multiple categories of objects.

Segmentation

Segmentation subdivides an image into several regions that might represent multiple objects. Usually, each pixel is assigned to exactly one region or the background (disjunctive segmentation). Segmentation into overlapping regions or hierarchical segmentation (distinguishing subobjects, such as compartments within cells) is considerably more complex [20].

Reaching human performance in automatic segmentation is hard. Fortunately, for many image mining tasks, approximate solutions are sufficient. For example, Ramm *et al.* [8] recognized cell bodies, neurites or debris. However, they did not attempt to match neurites with their respective cells (which would be very hard), but just used the ratio of their respective effective areas as their target indicator.

In this respect, one can distinguish between the following segmentation approaches:

(i) Marker-based segmentation. The area covered by a single marker or marker combination is taken as the spatial structure. If the assay is properly constructed, some filtering/smoothing and a thresholding operation (assigning each pixel above a certain threshold to the segmented area) are sufficient. For example, Camp *et al.* [21] used a fluorescently labeled tumor marker to locate tumor cells. Further membrane and nucleus markers, when co-located with the tumor marker, were taken as segmentations of the tumor cell membrane and nucleus.

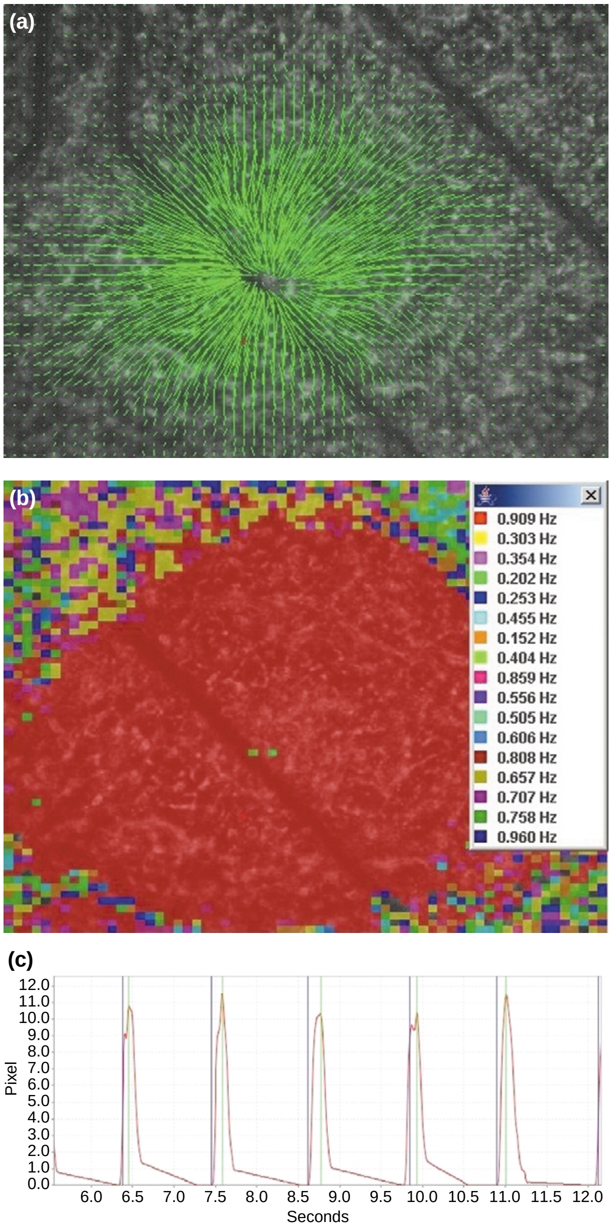
(ii) Object-based segmentation distinguishes individual objects, but does not accurately determine their boundaries. For example, nuclear regions can be taken as representatives of cells [22]. Object features can then be calculated over a fixed size area [23].

(iii) Contour-based segmentation. For shape-based or area-dependent features, the object contour must be determined to pixel precision. Such segmentation mechanisms are more elaborate and often gradient based. Membrane markers may assist segmentation [12].

Software for most automated microscope systems features proprietary segmentation methods with wide applicability [7,8,24].

Feature calculation

Information about objects in an image is condensed into features. Extensive sets of attributes can be calculated on a per cell basis. Features may provide quantitative measurements of color, local shape and texture (e.g. area/size or total fluorescence intensity per cell) [15]. Intensity can not only be calculated over the whole image, but also may be restricted to areas covered by another marker (marker combination, [24]). Derived features may indicate ratios



Drug Discovery Today

FIGURE 1

Snapshot of the motion analysis of beating cardiomyocytes. (a) Microscopy image of the cell aggregate overlaid with motion vectors at a particular point in time (to the next point in time, not shown). **(b)** The beating frequency calculated over the whole time series shown in color coding. **(c)** Elongation diagram of a particular pixel over multiple beats, quantifying the interval between the start and the peak of each beat. The frequency diagram shows the uniformity of rhythm, the motion vectors show that there is a spherical contraction and the peak quantification (summed up over a complete region) is used as a sensitive measure of activity. Microscopy images courtesy of Axiogenesis AG, Cologne, Germany.

of other features [12,21,25], such as the size of the nucleus relative to the cell. Based on a contour segment, shape properties such as the width and length of objects can be derived [8]. Features may also involve counting objects [26]. Zaidi *et al.* [27] calculated the spatial distribution of nuclear foci with the nuclei.

A huge number of more complex mathematical features can be calculated for individual regions/cells, and used to distinguish the spatial distribution or visible patterns/textures [28,29].

Objects or parts of the image can be tracked over time, and their motion and behavior analyzed (motion analysis) [17,18,30]. For example, aggregates of cardiomyocytes can be induced to beat in culture (Figure 1). A high-resolution camera observes the motion. Data analysis algorithms then derive quantitative measures of behavior (such as frequency and activity levels) that permit the precise characterization of cardiotoxic agents (H Bohlen, unpublished).

Most applications use a small hand-selected set of features. Applications can be tailored by offering only such a restricted feature set. Such a selection severely limits the expressive power of the application, however, and users frequently experience these limits. A huge offering of features, on the other hand, can be bewildering, with many of the more abstract features (e.g. texture properties) being hard to understand. Further research is therefore needed to determine how extensive feature sets can be presented and utilized.

Object classification

Objects and regions can be automatically classified into user-defined categories based on their features or properties (known or extracted from the image) [31]. This may require training of the system to automatically determine appropriate decision criteria. Although classification methods are widely applied to analyzing global patterns, they have just begun to be used in classifying individual objects in an image.

One can distinguish the following approaches to classification:

(i) Simple decision rules are defined by the user as a set of boundary conditions, depending on several calculated attributes. For example, Ramm *et al.* [8] classified cells according to size features into neurites (width 0.8–9.0 μm, length at least 30 μm), cell bodies (width and length at least 75 μm) and debris (any other objects). If the number of available attributes is high, finding the right set of boundaries is cumbersome.

(ii) Decision rules can also be set as a system of fuzzy rules that permit a gradual transition between different classes [20]. This is more accurate than simple rules, but even more complex to validate if there are many possible rules.

(iii) Clustering methods (unsupervised learning) distinguish groups of similar objects/images by employing a similarity measure defined on a set of features. A good clustering may be achieved if the number of expected groups is known and if a suitable set of attributes can be chosen [2,32].

(iv) Supervised learning methods, such as neural networks [33] or support vector machines [34], take a set of examples classified by the user and generalize the decision criteria to arbitrary feature values [2,28,29,33–36]. Huang and Murphy [37] compared an extensive set of features and different classifiers to correctly label spatial patterns created by proteins with different subcellular localization in HeLa cells. Selection of the right training samples, feature set and learning method is important for the success of this method. A certain level of experience with machine learning is required, so these methods are not yet generally available or applicable in this context.

Visual data mining

Once symbolic data are derived from images, the results can be analyzed using all data mining approaches. For example, Mitchison [38] clustered different drugs according to their dose-response profile. A dose-response profile was built as a vector of aggregated response features for each dose. The response feature itself was calculated from 93 features of 11 objects selected from an image. This makes it clear that selecting and deriving the right features requires some effort.

There is a growing understanding that, in addition to data mining algorithms, an integrated environment is needed that supports visual data mining [39]. Visual inspection of the analysis results in relation to the raw data and other symbolic information is necessary for several reasons: it is the basis for selecting features, objects and training samples; it provides an understanding of the capabilities of the automatic analysis; and it may trigger additional knowledge of the human expert.

According to Shneiderman, visual information seeking can be described as a focused sequence of tasks: “overview first, zoom and filter, then details on demand, view relationships” [40].

To support Shneiderman’s overview task and to provide a starting point for zooming and filtering, an image database needs to be tailored to the central aim of the application; this determines the structure of the overview and the objects most relevant to that question. For example, in high-content screening, the aim is often to prioritize among a set of genes, target proteins or lead compounds. In this case, the overview should start with a list of these objects, with any information highlighted that contributes to an object’s priority [41]. Such an image database could be described as data centric.

Other image databases are more image centric. In this case, the image stored is the most relevant top-level entity [42]. However, this approach only provides an anecdotal overview, as semantic relations need to be uncovered explicitly through targeted queries. Thus, such interfaces primarily support the image researcher. Image-centric databases may also provide content-based image retrieval (CBIR, whereby the user searches for images similar to the given one) [43]. Similarity, just as for clustering, is derived from image and object features.

A third category of image databases could be described as process centric, whereby the top-level view corresponds to laboratory or analysis processes. These systems primarily support inspection and quality control tasks.

Zooming and filtering operations in image databases need to operate both on raw images (image zooming) and on symbolic information (semantic filtering) [2,14]. The combination and integration of these two approaches makes image mining powerful. A further option for filtering in image databases is to define the notion of similarity between different objects by defining a similarity measure (distance) between vectors of numerical features calculated for each of the objects [14].

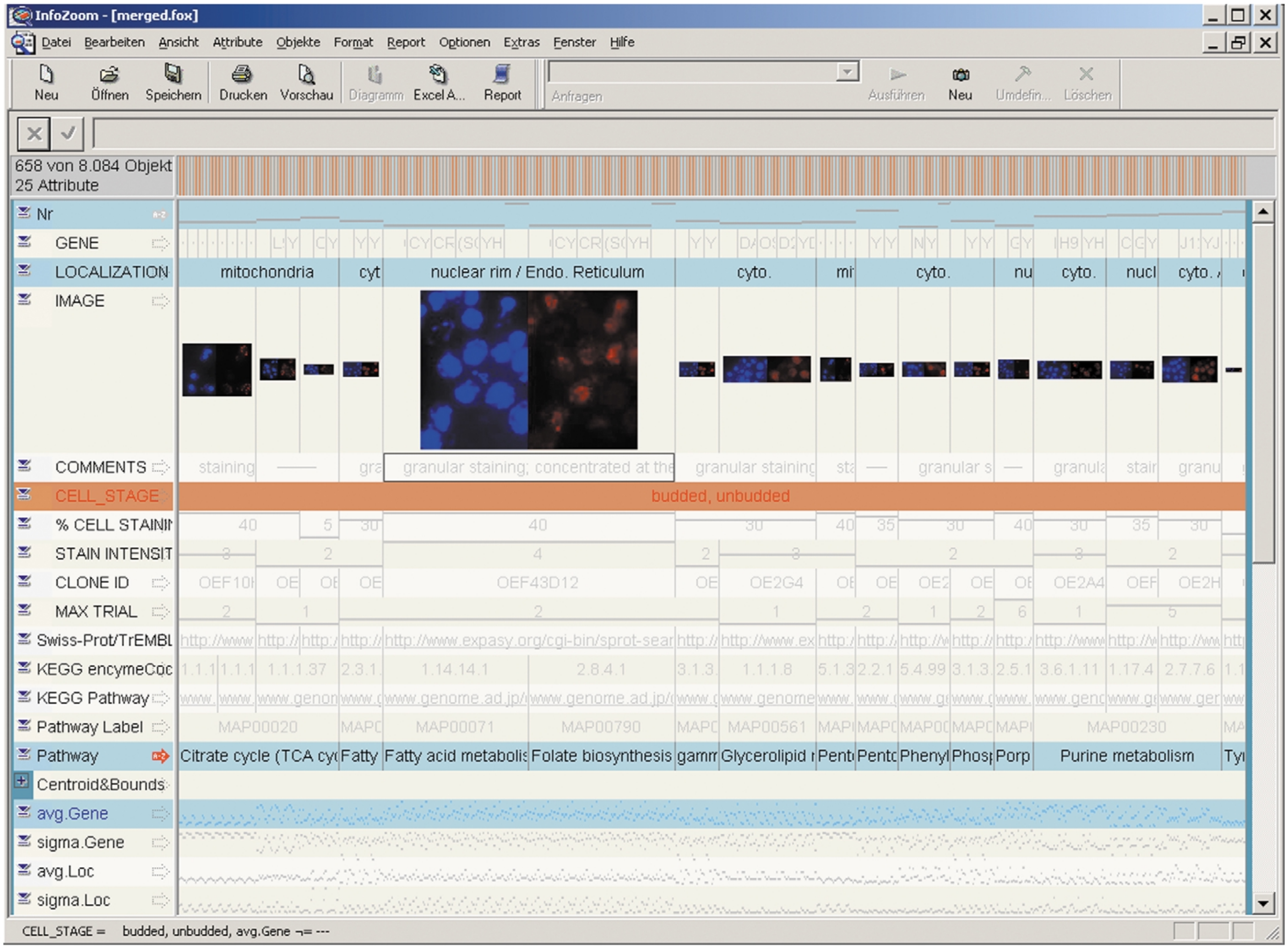
Traditional database queries are increasingly substituted by interactive methods of visual information seeking. The disadvantage of explicit queries is that the user needs to know what to ask for and will not notice if there is further information to be found under different keywords or headings. A hierarchical ‘drill down’ (details on demand, as in most file system browsers for example) can also become difficult with large data sets and multiple perspectives. The best options currently available combine information zooming [44,45] (Figure 2) for selection in multivariate descriptions with the highly flexible generation of views and diagrams for comparison [46,47].

Data management

Data management for image mining requires a combination of complex symbolic data models and large amounts of raw image data. The data model needs to capture a hierarchy of samples, experimental conditions, and objects within images, features and classifications. It is not surprising that these combined demands stretch data management approaches to the limit. On the other hand, complex transaction facilities are often not required for image databases, as information fragments are mainly added and rarely replaced. When raw and derived data serve as evidence of important conclusions, they need to be conserved in untouched form for scientific and regulatory reasons [2,42].

The first issue arising with image data management is the amount of storage needed for large images, such as video sequences. There is a trade-off between processing and storage demands. If the images can be analyzed in real time, only a fraction of the data needs to be stored permanently. For example, the Acumen Explorer [9] removes the image background while scanning in order to reduce storage demands. On the other hand, if all the images are stored, information can be extracted over night or later when needed, although the data management infrastructure will be more complex.

Another problem with data management for image databases is data modeling, that is, defining the entities and attributes needed for a particular experiment. Data modeling is required both for ancillary metadata, which is concerned with experimental setup [17,18], and for



Drug Discovery Today

FIGURE 2

Screenshot of the InfoZoom™ visual data mining tool — an effective and general implementation of Shneiderman's information-seeking paradigm. A selection of 658 records from a database of 8084 genes is shown [61]. Each column shows one record (gene), each row one attribute. The screen provides an overview of the relative frequency of values (e.g. the largest number of genes shown is related to the citrate cycle), because horizontally adjacent identical values are collapsed (the display is sorted according to the attribute 'pathway'). This is an example of an object-centric visualization. The user can 'zoom in' by clicking on any attribute value shown and thus display only those records. In this example, the user has already selected some values for the attribute 'cell_stage' (displayed as status in the lower-left corner). Records will occupy more space accordingly and image space will also enlarge. Zooming is not restricted to a hierarchical drill down. Attribute groups can be expanded on demand to show details. The tool can be coupled to various databases.

intrinsic metadata, which is related to the semantic content derived during the analysis. One can distinguish among the following approaches that are available.

It is possible to manage all information in flat files, both images (as raw data files) and symbolic information (in text, table or spreadsheet form). For example, Liebel *et al.* [23] used this approach to manage data in a screening system, employing a full-text search engine to retrieve data. The metadata (identification of the sample, etc.) are stored in the file name. This approach requires a lot of discipline and manual intervention for performing the analysis.

Some standard image formats already contain metadata inside the image file, for example, DICOM [48], MPEG-7 [17] and JPEG (EXIF data for digital cameras). These can be difficult to inspect and extend. If the metadata are stored outside the image (in accompanying files), they have to reference each other using an identification scheme.

Databases that manage shared data are not ideally suited to image mining applications. This is because performance is most often not optimized for storing large amounts of raw data, and the data model gets subdivided into many relations as the data get more complex and hierarchical.

In many cases, symbolic data are stored in a relational database, whereas the images are stored in a file system and are only referenced by the database. Consistency requires some discipline, as the information is managed in two largely separate entities. A typical example is an

Oracle database linked to images managed by the SRB (storage resource broker) [13]. The SRB locates and retrieves binary files within a computer network. Similarly, Manley *et al.* [5] use an Access database with the images stored in the file system. Commercial image databases, such as SIMS™ by SciMagix (www.scimagix.com) or IQbase™ by Media Cybernetics (www.mediacy.com), also permit the user to add metadata (beyond the keyword annotation of many simple image repositories), but still require integration efforts with data sources and analysis tools.

Complex data can be modeled more accurately in an object-oriented database, which is able to map even complex hierarchies. However, the lack of standards and the high demands imposed by the modeling have prevented a more widespread adoption. For example, Diallo *et al.* [42] used the Illustra object-relational database as the basis of their own framework. It enabled them to extend the set of data types handled by the database to geometric objects (such as three-dimensional regions of interest), with their own processing functionality (segmentation, visualization).

Recent developments try to manage data by integrating chunks of information in semi-structured form. Usually, each chunk can be described as an XML file, which is a modeling language that allows each file to carry its own complex structure. These approaches can very flexibly manage archives of heterogeneous data, but appropriate tools to visualize and mine these archives still have to be developed. An example of this category is the NeoCore database [49], a repository of mixed XML and raw data files. As the vast majority of computer applications (both open source and commercial) are beginning to incorporate this technology [50], the promise is that semantic web technologies [51] will, in the future, provide a high level of interoperability between different tools and applications [52].

For the time being, there remain several open issues regarding this approach. Because information resides in separate files, referencing to other pieces of information requires a location-independent identification scheme, as provided by the Life Science identifier proposal [53]. Furthermore, in addition to the syntactic interoperability provided by XML, there is an ongoing need to standardize higher level concepts and semantic descriptions [54].

Applications

Diagnostic support

Image-based diagnosis assigns samples or patients to particular categories, such as diseases or states. Often, decision criteria are fuzzy and there is considerable variability among human observers. Image mining may be used to derive and employ more consistent decision criteria. Image mining techniques can enhance diagnosis by calculating quantitative features or features not easily available to a human observer, and by suggesting and validating category boundaries, including learning categories from the data.

One of the prominent applications of diagnostic support is cervical cancer grading from microscopy images [33,55]. Applications include detection of rare cells in blood [35], cell classification [56] and, more generally, support for pathological diagnosis (e.g. using tissue microarrays [5]).

Diagnostic support can also be provided for medical imaging modalities, such as Alzheimer's diagnostics on PET (positron emission tomography) images [57], the classification of radiology images [58] or the use of photography to diagnose skin lesions [34].

Image mining can also be used to detect specific disease biomarkers. Schubert [19] developed a novel method to sequentially localize several different (in this case 17) antibodies, resulting in an expression vector of multiple proteins for each pixel. By clustering frequently appearing vectors (the space of potential patterns is much larger than the number of pixels in an image), this method could identify co-localization patterns (i.e. a set of proteins found in the same place) that specifically appeared only in the membrane of invasive immune cells.

A major barrier to the more widespread adoption of these techniques, however, is of a financial and legal nature. The pathologist or physician always has the final diagnostic decision. Automatic classification and diagnosis will thus be added cost unless they can show a significant advantage. Complex diagnostic procedures are expensive to validate and carry liability risks. Companies will therefore only invest in areas in which no suitable diagnostic procedure is available. However, automatic diagnostic procedures may increase diagnostic quality via enhanced reproducibility and by providing a second opinion.

Functional studies

Through systematic variation of experimental parameters and controlled observation of biological material, functional information can be derived and validated. In particular, the miniaturization and automation of cell-based assays enable the systematic study of biological systems. The analysis concentrates on the identification of functional relationships (between input parameters and results) and their statistical validation.

Zaidi *et al.* [27] studied mitotic partitioning in the nucleus by measuring the spatial distribution of nuclear foci and tissue-specific transcription factors throughout the cell cycle. Ortiz de Solórzano *et al.* [12] used a database of three-dimensional cellular images with semi-automatically segmented nuclei to quantify the tissue response to different intensities of radiation therapy in order to optimize treatment. An increasing number of groups are studying translocation phenomena [56], whereby the spatial distribution of molecules is captured through fluorescence imaging, quantified and studied in response to external factors. In all these applications, quantitative features involving spatial distribution are calculated, and plotted as data points to visually identify and verify dependencies among factors.

Three-dimensional and time lapse images are also of interest [16]. Such images have even been used to validate mathematical models of processes such as transport. For example, a quantitative model of Ran transport has been established and validated by time-lapse imaging [59].

Schubert [19] used a set of multiple markers and their co-localization to distinguish different phases of cell invasion, as characterized by distinct co-localization patterns common to all cells in a particular phase.

Image mining is still not widely used for functional studies, because equipment is expensive and varying experimental approaches demand considerable resources for setting up segmentation parameters, features and classifications each time.

Screening

The purpose of a screening experiment is the optimization of function. A large number of input parameters (e.g. compound libraries or experimental conditions) are varied and the best parameter set is selected that has the strongest positive effects measured by image features.

In contrast to methods that measure just a single quantity (such as flow cytometers), microscopic images of cells and tissues provide more information about spatial and temporal processes. Images also permit the distinction of multiple objects from different categories, such as cellular compartments, cell types or tissue regions. 'High-content screening' refers to processes that are defined spatially and temporally in the array of cells [7].

A large number of screening experiments have been described in the literature to assess factors that influence processes such as neurite outgrowth [8], gap junction blocking [24], centrosome duplication [22], tumor cell proliferation [60], nucleocytoplasmatic transport [59], Golgi integrity [23] or wound healing [25]. In all these cases, quantitative features are defined on a per object basis and then aggregated in a way that would not be possible on the whole image.

As particular assays can typically be re-used for different screens (the setup needs to be programmed only once) and are often less complex compared to functional studies, image mining is more cost effective here.

Challenges

Image mining comprises a bewildering array of methods and tools. Freely or commercially available applications need to balance between open tool sets that can only be operated by programmers, and closed applications that are too rigid in their usage or too expensive to develop and maintain. To effectively obtain practical applications, several advances are necessary:

(i) Data management needs to become more flexible and more easily distributed in Grid-like computer networks. The problem of interoperability and integration of multiple databases needs to be addressed. This will probably be achieved by enabling analysis and visualization tools to work on semi-structured data instead of coupling them to a rigid data model.

(ii) Image analysis needs to become end-user programmable. Training and self-learning on examples will make feature selection and specification of image analysis solutions much simpler than today.

(iii) Visual data mining tools must be able to be tailored to specific high-level tasks and workflows without programming a completely new system. The semantics of a screening task, for example, need to be reflected as guidance in the user interface.

Automated imaging equipment is available today, but software support still has a long way to go. Most applications in the literature are either realized through manufacturers' software or programmed in environments such as MATLAB, LabView, IDL, Visual Basic and others.

Conclusion

Image mining based on automated microscopy requires an image database that provides data management, image analysis and visual data mining. The main characteristic of such a system is a layer that represents objects within an image, and that represents a large spectrum of quantitative and semantic object features. To adapt the system to various experimental approaches, the user has to adopt a programming mentality. It is of equal importance to be able to interactively access the results quickly down to the image level.

References

1 Hsu, W. *et al.* (2002) Image mining: trends and developments. *J. Intell. Inf. Syst.* 19, 7–23

2 Chubb, C. *et al.* (2003) Semantic biological image management and analysis, Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03), IEEE, 69–76

3 Kapur, R. *et al.* (1999) Streamlining the drug discovery process by integrating miniaturization, high throughput screening, high content screening, and automation on the CellChip™ System. *Biomed Microdevices* 2, 99–109

4 Liu, C.L. *et al.* (2002) Software tools for high-throughput analysis and archiving of immunohistochemistry staining data obtained with tissue microarrays. *Am. J. Pathol.* 161, 1557–1565

5 Manley, S. *et al.* (2001) Relational database structure to manage high-density tissue microarray data and images for pathology studies focusing on clinical outcome. *Am. J. Pathol.* 159, 837–843

6 Beske, O. *et al.* (2004) A novel encoded particle technology that enables simultaneous interrogation of multiple cell types. *J. Biomol. Screen.* 9, 173–185

7 Abraham, V. *et al.* (2004) High content screening applied to large-scale cell biology. *Trends Biotechnol.* 22, 15–22

8 Ramm, P. *et al.* (2003) Automated screening of neurite outgrowth. *J. Biomol. Screen.* 8, 7–18

9 Grépin, C. *et al.* (2003) Improving quality: high content screening with the Acumen Explorer. *Current Drug Discovery*, 37-42.

10 Jäger, S. *et al.* (2003) A modular, fully integrated ultra-high-throughput screening system based on confocal fluorescence analysis techniques. *J. Biomol. Screen.* 8, 648–659

11 Carpenter, A.E. and Sabatini, D.M. (2004) Systematic genome-wide screens of gene function. *Nat. Rev. Genet.* 5, 11–22

12 Ortiz de Solórzano, C. *et al.* (2002) Applications of Quantitative Digital Image Analysis to Breast Cancer Research. *Microsc. Res. Tech.* 59, 119–127

13 Martone, M.E. *et al.* (2002) A cell centered database for electron tomographic data. *J. Struct. Biol.* 138, 145–155

14 Singh, A.K. *et al.* (2004) A distributed database for biomolecular images. *SIGMOD Record* 33, 65–71

15 Smeulders, A.W.M. *et al.* (2000) Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 1349–1380

16 Kam, Z. *et al.* (2001) Probing molecular processes in live cells by quantitative multidimensional microscopy. *Trends Cell Biol.* 11, 329–334

17 Rodríguez *et al.* (2005) Analysis and description of the semantic content of cell biological video. *Multimedia Tools and Applications* 25, 37–58.

18 Shotton, D.M. *et al.* (2000) Object tracking and event recognition in biological microscopy videos. Proceedings of the 15th International Conference on Pattern Recognition (ICPR 2000), 226–229

19 Schubert, W. (2002) Polymyositis, topological proteomics, technology and paradigm for cell invasion dynamics. *J. Theor. Med.* 4, 75–84

20 Biberthaler, P. *et al.* (2003) Evaluation of murine liver transmission electron micrographs by an innovative object-based quantitative image analysis system (Cellenger®). *Eur. J. Med. Res.* 8, 275–282

21 Camp, R.L. *et al.* (2002) Automated subcellular localization and quantification of protein expression in tissue microarrays. *Nat. Med.* 8, 1323–1327

22 Perlman, Z.E. *et al.* (2004) High-content screening and profiling of drug activity in an automated centrosome-duplication assay. *ChemBioChem* 5, 1–8

23 Liebel, U. *et al.* (2003) A microscope-based screening platform for large-scale functional protein analysis in intact cells. *FEBS Lett.* 554, 394–398

24 Li, Z. *et al.* (2003) Identification of gap junction blockers using automated fluorescence microscopy imaging. *J. Biomol. Screen.* 8, 489–499

25 Yarrow, J.C. (2004) A high-throughput cell migration assay using scratch wound healing, a comparison of image-based readout methods. *BMC Biotechnol.* 4, 21

26 Brinkmann, M. *et al.* (2002) New technologies for automated cell counting based on optical image analysis 'The Cellscreen'. *Cytotechnology* 38, 119–127

27 Zaidi, S.K. *et al.* (2003) Mitotic partitioning and selective reorganization of tissue-specific transcription factors in progeny cells. *Proc. Natl. Acad. Sci. U. S. A.* 100, 14852–14857

28 Huang, K. and Murphy, R.F. (2004) From quantitative microscopy to automated image understanding. *J. Biomed. Opt.* 9, 893–912

29 Murphy, R.F. *et al.* (2003) Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images. *J. VLSI Signal Process.* 35, 311–321

30 da Fontoura Costa, L. and Schubert, D. (2003) A framework for cell movement image analysis. Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03), 271–276

31 Duda, R. *et al.* (2000) *Pattern Classification*, Wiley

32 Kumar, A. *et al.* (2002) Subcellular localization of the yeast proteome. *Genes Dev.* 16, 707–719

33 Zheng, Q. *et al.* (2004) Direct neural network application for automated cell recognition. *Cytometry A* 57, 1–9

34 Maglogiannis, I.G. and Zafiropoulos, E.P. (2004) Characterization of digital medical images utilizing support vector machines. *BMC Med. Inform. Decis. Mak.* 4, 4

35 Price, J.H. *et al.* (2002) Advances in molecular labeling, high throughput imaging and machine intelligence portend powerful functional cellular biochemistry tools. *J. Cell. Biochem. Suppl.* 39, 194–210

36 Perner, P. *et al.* (2004) Mining images to find general forms of biological objects. Proceedings of the International Conference on Data Mining (ICDM`04), Springer, 60–68

37 Huang, K. and Murphy, R.F. (2004) Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinformatics* 5, 78

38 Mitchison, T.J. (2004) Small-molecule screening and profiling by using automated microscopy. *ChemBioChem* 7, 1–7

39 Keim, D. (2001) Visual exploration of large databases. *Commun. ACM* 44, 38–44

40 Shneiderman, B. (1996) The eyes have it: a task by data type taxonomy for information visualizations. In Proceedings of the IEEE Symposium on Visual Languages, IEEE, 336–343

41 Curtis, J. *et al.* (2004) Information management for entomology screening. *J. Biomol. Screen.* 9, 37–43

42 Diallo, B. *et al.* (1999) B-SPID: an object-relational database architecture to store, retrieve, and manipulate neuroimaging data. *Hum. Brain Mapp.* 7, 136–150

43 Müller, H. *et al.* (2004) A review of content-based image retrieval systems in medicine – clinical benefits and future directions. *Int. J. Med. Inform.* 73, 1–23

44 Spenke, M. (2001) Visualization and interactive analysis of blood parameters with InfoZoom. *Artif. Intell. Med.* 22, 159–172

45 Spenke, M. and Beilken, C. (2003) Visualization of trees as highly compressed tables with InfoZoom. In Proceedings 9th IEEE Symposium on Information Visualization, IEEE, 122–123

46 Saraiya, P. *et al.* (2004) An Evaluation of Microarray Visualization Tools for Biological Insight. In Proceedings 10th IEEE Symposium on Information Visualization, IEEE, 1–8

47 Perlman, Z.E. *et al.* (2004) Multidimensional drug profiling by automated microscopy. *Science* 306, 1194–1198

48 Güld, M.O. *et al.* (2002) Quality of DICOM header information for image categorization. *Proc SPIE* 4685, 280–287

49 Direen, H.G. and Jones, M.S. (2003) Knowledge management in bioinformatics. In Chaudhri, A.B. *et al.* (Eds.), XML Data Management, Addison Wesley.

50 Swedlow, J.R. *et al.* (2003) Informatics and quantitative analysis of biological images. *Science* 300, 100–102

51 Neumann, E. and Thomas, J. (2002) Knowledge assembly for the life sciences. *Drug Discov. Today* 7, S160–S163

52 Berman, J.J. et al. (2003) A tissue microarray data exchange specification: a community-based, open source tool for sharing tissue microarray data. *BMC Med. Inform. Decis. Mak.* 3, 5

53 Clark, T. *et al.* (2004) Globally distributed object identification for biological knowledgebases. *Brief. Bioinform.* 5, 59–70

54 De Roure, D. and Hendler, J.A. (2004) E-Science: the Grid and the Semantic Web. *IEEE Intell. Syst.* 19, 65–71

55 Mango, L. (1994) Computer-assisted cervical cancer screening using neural networks. *Cancer Lett.* 77, 155–162

56 Perner, P. *et al.* (2002) Mining knowledge for HEp-2 cell image classification. *Artif. Intell. Med.* 26, 161–173

57 Sayeed, A. *et al.* (2002) Diagnostic features of Alzheimer's disease extracted from PET sinograms. *Phys. Med. Biol.* 47, 137–148

58 Keysers, D. *et al.* (2003) Statistical framework for model-based image retrieval in medical applications. *J. Electron. Imaging* 12, 59–68

59 Smith, A.E. (2002) Systems analysis of Ran transport. *Science* 295, 488–491

60 Bhawe, K.M. *et al.* (2004) An automated image capture and quantitation approach to identify proteins affecting tumor cell proliferation. *J. Biomol. Screen.* 9, 216–222

61 Kumar, A. *et al.* (2000) TRIPLES: a Database of Gene Function in *S. cerevisiae*. *Nucleic Acids Res.* 28, 81–84