Classification of Yeast Cells from Image Features to Evaluate Pathogen Conditions

Peter van der Putten[a], Laura Bertens[a], Jinshuo Liu[a%],
Ferry Hagen[b], Teun Boekhout[b] and Fons J. Verbeek[a1]

[a]Imagery & Media Group, Section Imaging & BioInformatics,
Leiden Institute of Advanced Computer Science (LIACS),
Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
[b]Yeast Research, CBS Fungal Biodiversity Centre,
Netherlands Royal Academy of Science, Uppsalalaan 8, Utrecht, the Netherlands

ABSTRACT

Morphometrics from images, image analysis, may reveal differences between classes of objects present in the images. We have performed an image-features-based classification for the pathogenic yeast *Cryptococcus neoformans*. Building and analyzing image collections from the yeast under different environmental or genetic conditions may help to diagnose a new "unseen" situation. Diagnosis here means that retrieval of the relevant information from the image collection is at hand each time a new "sample" is presented. The basidiomycetous yeast *Cryptococcus neoformans* can cause infections such as meningitis or pneumonia. The presence of an extra-cellular capsule is known to be related to virulence. This paper reports on the approach towards developing classifiers for detecting potentially more or less virulent cells in a sample, i.e. an image, by using a range of features derived from the shape or density distribution. The classifier can henceforth be used for automating screening and annotating existing image collections. In addition we will present our methods for creating samples, collecting images, image preprocessing, identifying "yeast cells" and creating feature extraction from the images. We compare various expertise based and fully automated methods of feature selection and benchmark a range of classification algorithms and illustrate successful application to this particular domain.

Keywords: image analysis, yeast, image retrieval, *Cryptococcus neoformans*, feature selection, microscopy

1 INTRODUCTION

Yeast cells come in many appearances yet only few of them have been identified as being pathogenic. The virulence of the yeast cell can, in some cases, be derived from the morphology of the cell; in *Cryptococcus neoformans* thicker capsules are believed to be an indicator for virulence. In other words, virulence can potentially be measured from the morphology of the yeast cell. The aim of our study is to develop a measurement and classification system for the virulence of cryptococcal yeast cells using their morphological characteristics; the classification should allow class prediction of new unseen images. Classification is initially related to the image derived features but should be extended in multi-media like fashion to link to biochemical and genomic data.

In the next sections we will discuss the background of this study on the image analysis of the pathogenic yeast *Cryptococcus neoformans* by introducing aspects that we have identified as being important to this study. In general a classification system depends on the images that are fed into it; the better the image processing and analysis can be performed the better the classification system will become. To that end it is important to realize that image processing is greatly facilitated through the image preparation. Adding contrast to the specimen so that cells can be extracted is

[%] Current address: Computer School, Wuhan University, Wuhan, P.R.China
[1] Correspondence should be addressed to: Fons J. Verbeek, Imaging & BioInformatics, LIACS,
Leiden University, Niels Bohrweg 1, 2333 CA Leiden the Netherlands (fverbeek@liacs.nl; http://bio-imaging.liacs.nl)

crucial to the further handling of data. This paper therefore takes a rather holistic view on the construction of such a classification system.

Our study concerns the pathogenic yeast *Cryptococcus neoformans*. This basidiomycetous yeast can cause meningitis, meningoencephalitis, and pulmonary and skin infections. Infections occur mainly in immunocompromised patients, i.e. HIV-infected patients, transplantation patients and leukemia patients[4]. One of the most significant virulence factors of the fungus is the presence of an extra-cellular polysaccharide capsule[1, 5, 6, 13, 16]. Complementation of a capsule-deficient mutant clearly showed the relation between the presence of a capsule and cryptococcal virulence[5]. The thickness of the capsule can vary between strains, specific genetic constructs related to capsule biosynthesis, and between different environmental conditions[4, 6,24].

Although measuring the size and shape of the capsule seems straightforward, it has not often been applied. Using the morphology of the yeast cells, the obvious analysis is to look at the capsule thickness directly, either by automated or semi-automated methods[18]. Early attempts may have been hampered by the fact that the staining methods were improper for good image analysis in that the staining results were not reproducible (cf. §2.2). We have experienced such as well in earlier work and newer staining methods have opened possibilities for large scale analysis[17].
Applying an image analysis driven method will allow deriving more features than just the capsule thickness. Rivera et al.[18] have analyzed samples of mouse brain and lung infected with *C. neoformans* to look at capsule thickness and cell volume. The cells were segmented from the images using hand-tracing. The features were derived by estimating the radius and computing the features with the analytical equations of circle area and sphere volume. The features were expressed in SI units.

A true automated tool has not been presented to date. Given the relatively simple shapes of the cryptococcal cells an effort to develop such a system should be undertaken. Extracting features from cells should not be restricted to the capsule thickness but extend to a broader range of features that can be extracted from images of cryptococcal cells. Features should inform us about the different classes of cryptococcal cells that can be distinguished. Ideally, a sample is taken from a population, i.e. a yeast culture, and from the features a distribution of the virulence can be found leading to a further understanding of the virulent state of a particular culture.

Performing image analysis and collecting measurements will allow learning of the size and shape of the yeast cells. This learning should be established in classifiers so that new images can be understood. Image analysis tools allow extracting lots of features; not just those related to the capsule thickness. Therefore we will investigate, in a limited set of features, which other features are relevant in such a classification system and for that matter will contribute to the predictive value of this measurement system. These features are related to the information in the image, the density distribution as well as the shape of the yeast cell and therefore these features are expressed in the pixel-image function domain rather than SI units.

The approach of image analysis and classification is relatively new in the field of yeast genomics; it is a very necessary approach though, as such systems will in the near future be used in large scale screens. This paper presents the first steps towards such a system with a focus on the image derived features. Features and images should be stored in a database and interoperable screening will allow retrieving other features related to the same sample from other databases. This is the typical trend currently seen in the bioinformatics research fields, i.e. an integrative approach of bioinformatics retrieving information from a broad panel of related bio-medical, molecular and organismal databases. All considered, this paper presents a holistic description of end to end process from growing the yeast, capturing images, image segmentation, feature extraction and classification.

The structure of this paper is as follows. Section 2 describes the materials and methods; this is a broad collection of facts about the biology, the imaging as well as the computing aspects. The rationale of the structure of this section is to get a good overview of the processes at hand in the experiments described. Section 3 describes the results that were obtained for the computing related issues; the image processing and the classification. In section 4 the perspective is given of the applications of the system that we present. Finally, in section 5 we summarize by presenting our conclusions and providing insight in future work.

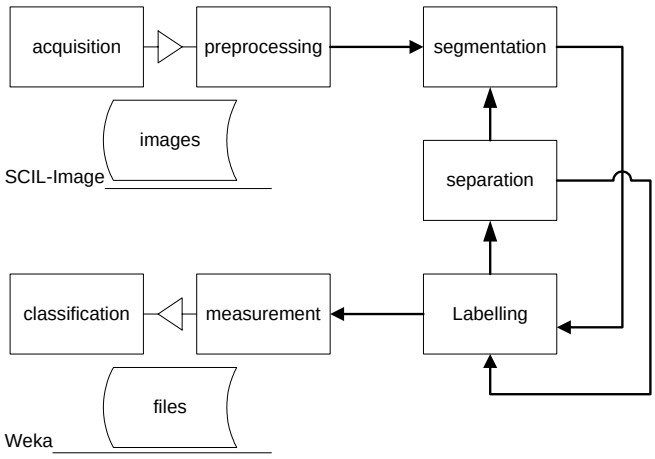


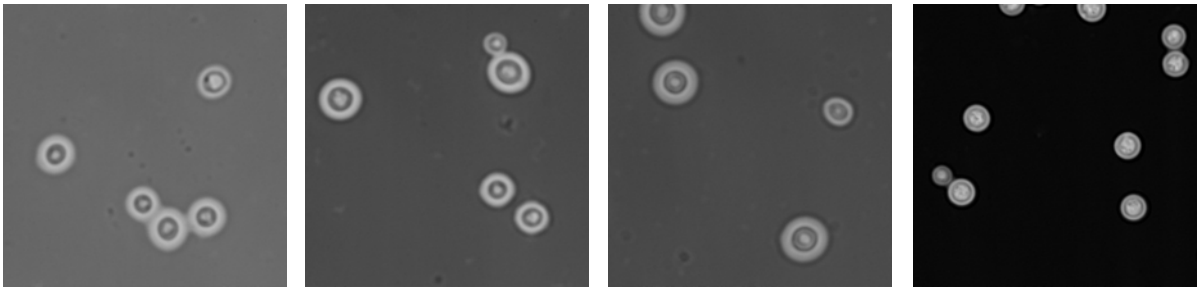
Figure 1. The system process flow. The thick black lines indicate the flow of the data as they are passed to another module. The open arrows indicate the flow of files (image in and measurement file out) as used in the system. ScilImage is used for all image related work (cf. § 2.4-2.5); Weka is used for work related to the classification and clustering (cf. § 2.6)

2 MATERIAL AND METHODS

The pipeline for experimental data consists of a range of modules passing files and/or data. At the onset of the pipeline is the acquisition. The images first need to be enhanced and processed in order to be suitable for a segmentation procedure. The segmentation procedure consists of the segmentation itself so that binary images can be evaluated in terms of shape analysis. Each cell is evaluated separately and therefore segments need to be established as such and then processed such that individual cells are extracted from the segments and measured. The results of the measurements are written to tab-delimited files (TDF) which can be imported in any other module.
Two data-types are distinguished, the images that are transferred over the image processing modules and the TDF data that are imported in the classification modules. The processes that are distinguished in the pipeline are incorporated in seven modules. The processing pipeline is illustrated in Figure 1.

2.1 Producing the yeast strains

Preliminary investigations were made using a variety of *Cryptococcus* strains from the collection of the CBS Fungal Biodiversity Centre using some media that are known to influence capsule size. Among these were Littman medium[15], Golubev medium[7] and the recently described Sabouraud media with or without MOPS, HEPES, pH5.5 and 7.3, and 1/10 diluted Sabouraud medium[25]. Unfortunately, results obtained using these growing conditions were not optimal. Therefore, we decided to use Potato Dextrose Agar (PDA: 230 ml potato extract, 20 g dextrose, 15 g agar, 770 ml water, pH 6.6) that according to our experiences at CBS result in highly mucoid colonies and capsulated yeast cells in many basidiomycete yeast species. Two *C. neoformans* strains were used in the final analysis, namely an acapsular mutant CBS 7926 (Cap 59- mutant of NIH B-3501, E.S. Jacobson) and a capsule containing isolate CBS 7936[#] (Cap 67- mutant of NIH B-3501, E.S. Jacobsen). The strains were maintained in the gas phase above liquid nitrogen at -135ºC, subcultured twice at PDA (48 h at 25ºC) and investigated for capsule size. In addition CBS 6955 (=ATCC 32608 = NIH 191), a *Cryptococcus gattii* strain, was used in our experiments. This strain was cultured on the standard YPGA medium (1% yeast-extract, 1% peptone, 2% glucose, 2% technical agar #3, all in w/v); the staining procedure for this strain was equal to the other two strains in this study.
In order to enhance the contrast for the imaging, we used the nigrosine staining method for our experiments. All yeast cells (CBS 7926, CBS 7936 & CBS 6955) were stained in 9 µl of 5% (w/v) nigrosine in water. The yeast cultures were inspected under the microscope to inspect if the culture was indeed according to expectations. Growth of the yeast culture on agar-plates was tested with a Leica Stereo microscope. Cultures that were proven to be of good quality were used in the experiments and slide preparations were obtained.



A **B** **C** **D**

Figure 2. Four examples of samples taken from a yeast culture negatively stained with nigrosine (cf. section 2). Each image is a sample and from each culture at least 20 images are obtained through random selection over the slide. The operator criterion is that in a selected field of view sufficient cells are present. The samples are by no means taken to establish the numerical density of the yeast in the specimen preparation.
Figure 2A depicts cluttered cells that need be separated in the preprocessing phase so that each cell can be quantified.
Figure 2B illustrates a budding yeast cell. The "new" cell is excluded from the analysis. In the preprocessing the bud is separated from the parent cell and the parent cell is used in the analysis.
Figure 2C depicts a cell that is captured incompletely. Cells on the image border are excluded from analysis.
Figure 2D depicts another yeast culture with cluttered and budding cells as well as cells on the image border.

2.2 Notes on Staining

In the image acquisition phase the focus was to obtain images of sufficient quality for images analysis. The starting point for this study was to use 2D images. From earlier research with cryptococcal cells we have learned about the quality of the staining; the staining methods that we have focused on are the so called background staining techniques that do not stain the specimen but rather enhance the microscopy features by making the background more light dense. Traditionally, yeast biologists were using Indian Ink in the microscopy preparation, the apparent disadvantage of this staining method in image analysis and specimen classification is that it is hard to get a reproducible staining; moreover application of this staining results in blurring of capsule margins. With the nigrosine staining method we were able to get reproducible staining with little artifacts. At the same time, and of equal importance to the reproducibility, the staining method rendered an excellent quality of the visualization of the yeast cell and its capsule. Moreover, it can be applied as a standard procedure and therefore it is easily included in the standard workflow in a yeast biology laboratory. Some examples of images that are obtained with the nigrosine staining are depicted in Figure 2. With the current protocol for specimen preparation and image acquisition images with very little artifact and noise are obtained.

2.3 Image acquisition

The slides were prepared for image acquisition on a Zeiss Axioskop with a PlanApoChromat 63x (NA 1.40) oil lens. The image acquisition was realised with an Adimec MXI2P black and white CCD camera mounted on the Zeiss Axioskop (Zeiss, The Netherlands) and connected to a Pentium 3; the acquisition was controlled by the Research Assistant (vs3.) software. The Adimec MXI2P acquires images with a dynamic range of 8 bits; the images are sized 640x444 pixels and stored as tiff files.

2.4 Image preprocessing and segmentation

The image processing is crucial to the experimental set up as the image processing pipeline generates the results for the final classification procedures. In itself the image processing falls apart in three discrete phases, namely the image acquisition, the preprocessing and the image analysis (feature extraction). The processing and analysis of the images is completed within the SCIL-Image environment[19]. This image processing environment is an extensible software package that fits well for scientific research. To complete the tasks of the research that is described in this paper a range of new routines were added to the package.

The preprocessing step consists in preparing the images for a segmentation procedure so that each individual yeast cell in the image is measured through a number of features. Ideally the yeast cells are evenly distributed over the microscope slide and the yeast cells captured in images are nicely separated. This is, however, not always the case and it can not be

imposed (cf. Fig. 2A-D). In the images we find budding cells, i.e., cells in the process of division, and cells that are cluttered together as well as dead cells. The classification of yeast cells should be based on features that are derived from single cells. The effort in the preprocessing and segmentation step is to accomplish that particular goal. This goal is achieved by applying firstly the appropriate filters (cf. Figure 3B & 3C); secondly performing the segmentation and thirdly processing each segmented image in such a way that only features from single cells are extracted. This requires application of heuristics in the segmentation process. An image is successfully processed if each of the cells that are completely visible in the image is transmitted to the image analysis part and its shape can be measured.

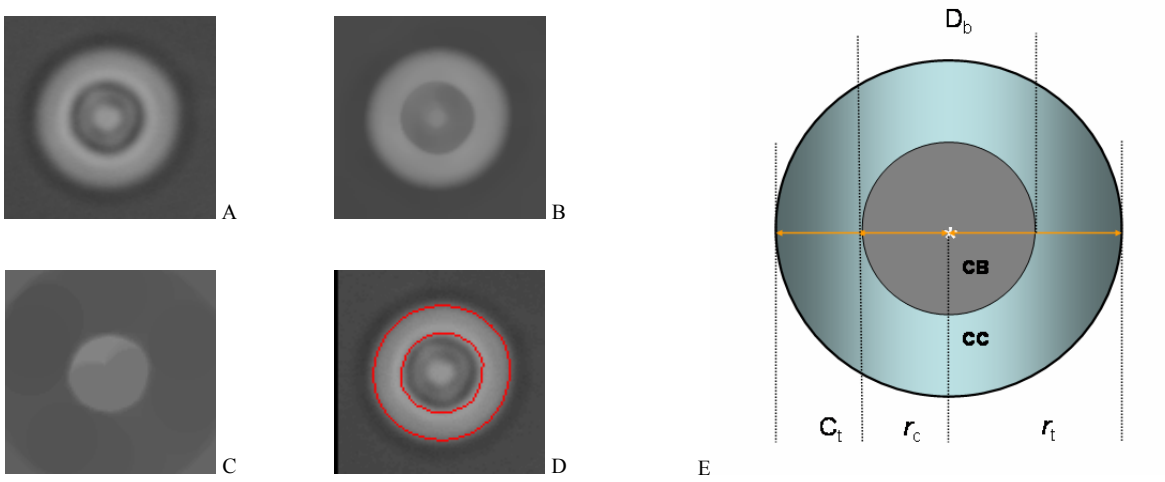


Figure 3. Imaging of the yeast cell with nigrosine background staining and showing obvious features from the cell.
Figure 3A (panel, top left), depicts one cell with a dark background. The complete cell consists of the capsule and the cell body. The capsule is less light dense and thus using bright-field microscopy appears as white in the image of a yeast cell. The cell depicted has a relatively large capsule. The next three panels illustrate the steps to segmentation and the result.
Figure 3B shows the result after filtering and enhancing the capsule so that through segmentation the total cell can be extracted.
Figure 3C shows the result of filtering and enhancing the cell body of the yeast cell. Segmentation and "XOR" on the binary images produces the required result of a separate measure for the cell body and the capsule.
Figure 3D shows the result as superimposed on the original image.
Figure 3E (left panel), depicts a model of the yeast cell with a capsule. Using this schematic drawing the initial features for the recognition of the pathogen yeast cell can be understood. These features are closest to the recognition of the biologist. Thickness of capsule and cell radius are illustrated clearly. D_b = diameter cell body, C_t = capsule thickness (in pixel units), r_c = radius of cell body, r_t = radius of total cell. CB= Cell body, CC = Capsule. CB and CC are also used for surface area of cell body and capsule respectively.

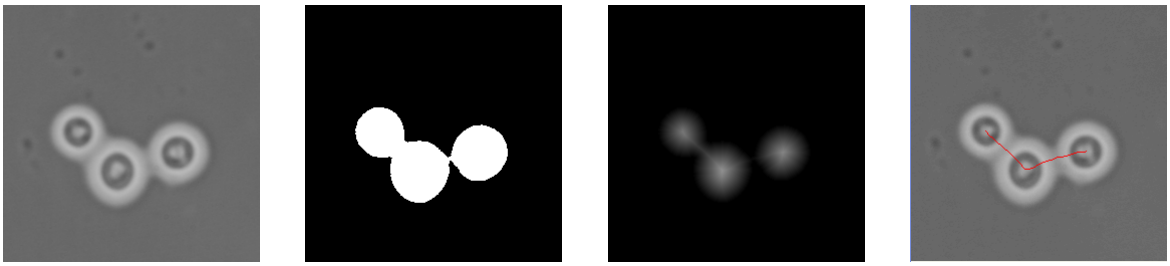
Figure 3E is supportive in understanding the segmentation procedure; a cryptococcal yeast cell consists of a cell body and a thick (or no) capsule. The segmentation is a multi-layered process as we have to be sure to extract the individual cells in the right way. Following, we list the processes involved in the segmentation procedure.

Preprocessing

The segmentation starts with a straightforward bi-level threshold operation which results in a binary image with just the capsules. In this phase of the segmentation it is, however, important that the outer contour of the cell, i.e., the outer boundary of the capsule is detected accurately. Using propagation a mask is created over the area that contains the entire cell. Next, of all objects in the image, the objects that touch the boundary are established and by an XOR operation these are excluded from the binary image. In doing so, the objects on the boundary, often incomplete shapes, do not contribute to the measurements (cf. Fig. 2C & 2D). The small sized objects are removed.

Labeling

The next step is a labeling operation so that we can extract and address each of the cells present in the image. In this phase we have to evaluate whether or not the cells are cluttered. For this evaluation the characteristic that intact individual cryptococcal cells are circular is used; this is accomplished with a circularity criterion which approximates one (1.0) for a circle. Cells that measure as circular are further segmented for measurement and cells with an aberrant circularity are processed in the separation module (cf. Fig. 1).



A **B** **C** **D**

Figure 4. The processing pipeline for the segmentation and segmentation of cluttered cells.
Figure 4A. A group of touching yeast cells as found in Figure 2A.
Figure 4B. The result of the segmentation process. The labeled component would not adhere to the heuristic of circularity and hence it would be considered a cluster that needs to be separated.
Figure 4C. The distance transform of 3B showing distinct maxima at the centroids of the individual cells.
Figure 4D. The skeleton, derived from the distance transform and superimposed on the original image (black line).

Segmentation

The circular mask is used to extract each of the labeled cells and perform a precise segmentation of the capsule (cf. CC in Fig 3E). This is completed in a buffer image of just size of the bounding box of the shape. The circular mask, obtained from the labeling is used in an XOR operation to find the area of the cell body (cf. CB in Fig 3E). After segmentation we have obtained two new masks, namely one for the capsule (CC) and one for the cell body (CB). In Figure 3D a segmentation result is depicted by superimposing the contours of CC on the original image. The masks and the buffer image are passed to the measurement module.

Separation

In case the circularity is not approximating 1.0, it is probable that the label represents a clutter of cells. An example of such a clutter is depicted in Figure 4A and in Figure 4B the binary label is shown. These cells need to be separated in order to be able to use them in the measure module. To that end, watershed segmentation is performed by applying a distance transform on the binary image of the clutter (cf. Fig. 4C) and from the distance image a non-branching skeleton is derived (cf. Fig. 4D). Superimposing the skeleton on the clutter helps to extract the maxima. Passing from one maximum to the next detects the minimum where the cells should be separated. The separation produces a (filled) mask for each of the cells in the clutter and the process continues with a labeling (cf. Fig. 1). The correctly separated cells are now processed by the segmentation module and prepared for measurement.

The individual cells that are probed in the image analysis phase are also used to determine the training/test set for the later classification. There is a range of features that can be measured from the shape of the cell and more specifically the capsule of the cell. In routine practice, the yeast biologist will evaluate capsule thickness through the microscope and possibly relate that to the radius of the cell. From digitized images, however, much more features can be derived. We will apply the measurements on the results from the segmentation procedure which are: a buffer image with the density image of one cell (cf. Fig. 3A) and two masks, i.e., one for the cell body (CB) and one for the capsule (CC).

2.5 Image analysis and feature extraction

Initial analysis is directed to the features the biologist will check when examining a sample of a yeast culture. As indicated in the introduction, the thickness of the capsule may be an indication of the level of virulence of a particular yeast isolate. Therefore, this feature needs to be analyzed in a reproducible manner. As capsule size is easily made objective by digital measurement, relative measures are computed comparable to what the yeast biologist does routinely by investigating yeast cells by bright field microscopy using negatively stained cells. In Figure 3E a schematic drawing is given to illustrate the measurement of the capsule thickness.

We have taken the image moments as a starting point of our analysis and as the objects are relatively simple shapes these moments provide sufficient information for a classification on the basis of shape. From the image moments a set of features is derived that is used in the classifications and clustering.

In a two-dimensional density image, the image moments are expressed as: $m_{pq} = \iint x^p y^q f(x,y)\, dxdy$ (1),

where p+q indicates the order of the moment.

For the case of a sampled image of size NxM this translates to:

$$M_{pq} = \sum_{x=0}^{N-1}\sum_{y=0}^{M-1} x^p y^q f(x,y) \quad (2)$$

One should realize that in case of images with binary objects the function f(x,y) filters all irrelevant image information. In the case of binary images the image moments provide information on the geometrical distribution of a point set.

The moments are made translation invariant by centering on the mean of the distribution. The mean is computed from the zero and first order moments. The centralized moments are expressed as:

$$\mu_{pq} = \sum_{x=0}^{N-1}\sum_{y=0}^{M-1} (x-\bar{x})^p (y-\bar{y})^q f(x,y) \quad (3)$$

The first order moment in a binary image equals the object area; ergo the μ_{00} of CB is the area (in pixels) taken by CB and μ_{00} of CC is the area of the capsule. The sum of these zero order moments is the total area of the cell and thus we can express the relative area of CB and CC in terms of zero order moments as:

$$A_{CB}^{rel} = \frac{\mu_{00}^{CB}}{\mu_{00}^{CB} + \mu_{00}^{CC}} \quad (4a) \qquad \text{and likewise,} \qquad A_{CC}^{rel} = \frac{\mu_{00}^{CC}}{\mu_{00}^{CB} + \mu_{00}^{CC}} \quad (4b)$$

Instead of deriving r_t and r_c (cf. Fig. 3E) from the area by using the analytical equation of a circle we use the data to find radii in the shape. These are the semi-major and semi-minor axis of the distribution, also known as the moments of inertia. The semi-major/minor axes are computed[21,22] from the centralized second order moments as:

$$\alpha = \left(\frac{2\left[\mu_{20} + \mu_{02} + \sqrt{(\mu_{20}-\mu_{02})^2 + 4\mu_{11}^2}\right]}{\mu_{00}} \right)^{1/2} \quad (5a), \text{ and } \quad \beta = \left(\frac{2\left[\mu_{20} + \mu_{02} - \sqrt{(\mu_{20}-\mu_{02})^2 + 4\mu_{11}^2}\right]}{\mu_{00}} \right)^{1/2} \quad (5b)$$

In the same manner the relative thickness is derived from the semi-major/minor axis:

$$T_{CC}^{rel-msjor} = \frac{\alpha^{CC}}{\alpha^{CB} + \alpha^{CC}} \quad (6a) \qquad \text{and likewise,} \qquad T_{CC}^{rel-minor} = \frac{\beta^{CC}}{\beta^{CB} + \beta^{CC}} \quad (6b)$$

Beside the major and minor axis we can derive a radius of gyration in both the x and the y direction:

$$\gamma_x^{CC} = \sqrt{\frac{\mu_{20}}{\mu_{00}}} \quad (7a) \text{ and likewise, } \gamma_y^{CC} = \sqrt{\frac{\mu_{02}}{\mu_{00}}} \quad (7b), \text{ and additional unified form } \gamma_{xy}^{CC} = \sqrt{\frac{\mu_{20}+\mu_{02}}{\mu_{00}}} \quad (7c)$$

The third order moments relate to the skewness of the distribution[21] in both the x and y direction as follows:

$$Sk_x^{CC} = \frac{\mu_{30}}{\sqrt[3]{\mu_{20}^2}} \quad (8a) \qquad \text{and likewise,} \qquad Sk_y^{CC} = \frac{\mu_{03}}{\sqrt[3]{\mu_{02}^2}} \quad (8b)$$

The kurtosis (peakedness) is derived from the fourth order moments in both the x and y direction as follows:

$$K_x^{CC} = \frac{\mu_{40}}{\mu_{20}^2} - 3 \quad (9a) \qquad \text{and likewise,} \qquad K_y^{CC} = \frac{\mu_{04}}{\mu_{02}^2} - 3 \quad (9b)$$

From the moments a set of 7 invariants can be derived[8, 12]. These invariants are computed through a normalization of the centralized moments. We have employed the first four invariants; we will express the invariants in terms of normalized moments without further addressing the normalization step. The first and second invariant are derived from second order moments and computed as:

$$\phi_1 = \eta_{20} + \eta_{02} \quad (10) \qquad \text{and,} \qquad \phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11} \quad (11)$$

The third and fourth invariants are computed from third order moments as:

$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \quad (12) \qquad \text{and,} \qquad \phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \quad (13)$$

In our experiments the moments are computed for both the binary image and the grey-value images. The binary images are the cell body (CB) and the cell capsule (CC), both available from the segmentation. In addition, the binary images are used to mask out the grey-values under the area of CB and CC respectively, so that these can be used to compute a grey-value moment set. The equations 4-13 provide a lot of features that require computation of the centralized moment-set (eq. 3), thus the moment sets are first transposed to the centralized form. This is accomplished for CB in binary and grey-value images as well as CC in binary and grey-value images. The binary measurements are related to the geometrical distribution of the shape whereas the grey-value measurements, in the way applied in our experiments, are related to the density distribution under the shape. All features are derived from the centralized moments and formulated for the CC; in all cases the feature is *mutatis mutandis* derived for the CB. In addition, the relative area and relative thickness have been introduced to rule out the effect of the size of the cell. The relative area (Eqs. 4a,b) is derived from the zero order moment (in the binary case equal to area) and the relative thickness (Eqs. 6a,b) is derived from the semi-major and the semi-minor axis (cf. Eqs. 5a,b).

2.6 Clustering and Classification Experiments

The clustering and classification is completed with the Waikato Environment for Knowledge Analysis (WEKA)[23] which incorporates a wide variety of pattern recognition methodologies. Different methods are easily compared and data import is dealt with through standard file formats. Feature selection, clustering and classification modules from WEKA were used to complete this study.

As stated before, the goal from a biological perspective is to classify yeast cells into potentially virulent and non-virulent classes. We approximated this objective by building classifiers that can distinguish between different classes of yeast cells that resemble either virulent or non-virulent classes, with respectively thick or thin capsules. The image dataset that we have used to conduct our experiments is sufficient in a prove of concept setting.

A wide variety of clustering and classification models have been built to investigate the usefulness of the various features generated by the image preprocessing phase, ranging from features that measure capsule size directly to the more abstract moment invariants. For completeness we also experimented with various types of classification algorithms, though we feel that at this stage this is not a major factor in the quality of the final classifier.

A major point we would like to mention is that distinguishing between images may not be the same as distinguishing between categories of yeast cells. For instance, differences can be caused by varying environmental conditions when growing the cells, or can result from staining, image acquisition, feature extraction and other preprocessing steps. It is crucial to control these conditions as much as possible, which is only achievable to a certain extent and the rise of collaborative web collections of images may make matters rather worse than better, as image production becomes more separated from image distribution and analysis and collections become heterogeneous. So, in addition, it is essential to focus on feature extractors that extract the right type of information, i.e. focusing on cell characteristics, and this was a core issue we have kept in mind for our entire approach.

Three sets of images were available; a *Cryptococcus neoformans* strain with thick capsules (7936), a *Cryptococcus neoformans* mutant with thin capsules (7926) and a related strain, namely *Cryptococcus gattii* (6955) also with a thick capsule. This allows us to zoom in on detecting cell characteristics that are typically associated with virulence. The third set of images comes from a different strain (6955) which allowed us to check the performance of classifiers if more classes are introduced. We performed a univariate analysis to estimate the predictive power of each attribute, using the information gain measure. To explore the usefulness of the set of attributes from a multivariate point of view we performed a clustering on the data sets and evaluated the mapping of classes to resulted clusters. We then created an array of classification models using the procedure outlined below. All these experiments were carried out for the two-class (7926 vs. 7936) and three-class problem (7926, 7936, 6955).

As discussed the image preprocessing procedure produces an instance for each cell, resulting in a number of instances for each of the classes 7926, 7936 and 6955. A number of features produced by the preprocessing procedure were deemed either irrelevant for classification (f.e. x,y position), not invariant (f.e. size/area/radius) or prone to measuring differences in imaging or preprocessing conditions rather than cell characteristics (f.e. distributional characteristics of the inner part of the cell; same for binary images of the capsule). The remaining set of attributes, the base set, included gyration (x, y, unified for binary and gray values, cf., Eqs. 7a,b,c); variance, skewness (cf. Eqs. 8a,b) and kurtosis (cf. Eqs. 9a,b) of the distribution of the capsule; the first and second moment invariants (cf. Eqs. 10,11; binary and gray);

surface inner area compared to entire cell (cf. Eqs. 4a,b; binary and grey) and thickness of capsule compared to entire cell (cf. Eqs. 6a,b; binary). To investigate the contribution of various attributes, classification models were built on ten different subsets of attributes, see the results section for details.

As discussed both clustering and classification experiments were carried out. For the clustering experiments we used standard k-means clustering with two, respectively, three clusters and created a matrix to compare the distribution of classes over these clusters. The classification algorithms applied were decision stumps (split on single attribute), J48/C45 decision trees, naive Bayes, 1-nearest neighbor and 5-nearest neighbor. For the latter three algorithms we investigated two variants, one using all attributes available, and one based on selecting the most important attributes first. Attribute selection was performed on train sets only using the correlation based feature subset method (CFS) with best first forward search[9, 10, 23]. Classifiers were evaluated based on ten runs of ten fold cross validation for each classification algorithm – attribute set combination.

Table **1:** Predictive power of the attributes for the two-class and three-class classification tasks, measured in information gain on the full data set (15 top predictors out of 50). The abbreviations of Fig. 3E are used to indicate the measurement at hand; a suffix of "b" is added if it concerns measurement of a binary object, whereas a suffix of "g" is added if the measurement concerns a gray-value object. The corresponding Eq. cf. §2.5 is given in the second column.

Two class problem (7936;7936)		
Attribute	**Eq.**	**InfoGain**
φ1 CC-g	10	0.91
φ1 CB-b	10	0.91
φ2 CB-g	11	0.91
relative area CC-g	4b	0.87
φ1 CC-b	10	0.87
rel. area CC-b	4b	0.87
rel. thickness α CC-b	6a	0.87
rel. thickness β CC-b	6b	0.87
kurtosis y CC-g	9b	0.80
kurtosis x CC-g	9a	0.75
φ3 CB-g	12	0.70
φ4 CB-g	13	0.66
gyration ratio CC-b	7c	0.56
gyration x CC-b	7a	0.56
gyration y CC-b	7b	0.56

Three class problem (6955;7936;7926)		
Attribute	**Eq.**	**InfoGain**
φ1 CB-g	10	1.23
φ1 CC-g	10	1.00
φ2 CB-g	11	1.00
φ1 CC-b	10	0.97
relative area CB-b	4a	0.97
rel. thickness α CC-b	6a	0.97
rel. thickness β CC-b	6b	0.97
rel. area CC-g	4b	0.94
kurtosis y CC-g	9b	0.86
φ3 CB-g	12	0.81
φ4 CB-g	13	0.77
kurtosis x CC-g	9a	0.68
skewness y CC-g	8b	0.56
gyration y CB-g	7b	0.55
gyration ratio CC-b	7c	0.52

3 RESULTS

In this section we present the results of our experiments. Without sufficient results in the image processing modules no classification would be possible, therefore, we first summarize the results obtained through preprocessing and segmentation. A number of clustering and classification experiments are summarized in tables 2-4; the specific focus of the classifications and the content of the Tables is discussed in the second part of this section.

3.1 Preprocessing and segmentation

For the experiments described in this paper in total 84 images were processed. The images contained fully separated as well as cluttered cells (cf. §2.4, Fig. 3A-D and Fig. 4A-D). For the CBS 6955 cells 18 images were processed and 75 cells were extracted, for CBS 7926 cells 30 images were processed and 136 cells were extracted, whereas for CBS 7936 cells 36 images were processed and 66 cells were extracted. The difference in numbers in the CBS 7936 and CBS 7926

is caused by the fact that CBS 7926 is a mutant strain with significantly smaller size as they practically do note have a capsule; apparently more of these cells were in one sample.

For each of the image sets (6955, 7926, and 7936) the results of cells that were successfully segmented and measured were saved to TDF files. These files were imported in the classification environment. The files contain a large number of features some of which are not relevant for the classifications we have pursued in our experiments. A selection was made on the basis of info-gain measurements of the features.

Table **2:** Allocation of classes over clusters for two class – two cluster, three class – three cluster and three class – two cluster experiments

	Cluster 1	**Cluster 2**
7926	0	136
7936	66	0

	Cluster 1	**Cluster 2**	**Cluster 3**
7926	130	0	6
7936	0	64	2
6955	0	28	47

	Cluster 1	**Cluster 2**
7926	134	2
7936	0	66
6955	0	75

3.2 Clustering and classification results

Here we present the results for the various data mining experiments. First, we assessed the predictive power of individual attributes by calculating the information gain over the full training data, see Table 1 (15 top predictors out of 50). It is interesting to note the dominance of grayscale over binary image attributes. There is no clear winner between inner area and capsule attributes. The first and second moments invariants (cf. Eqs. 10, 11) dominate the top predictors.

The results of the clustering experiments can be found in Table 2. If clusters emerge that have a natural mapping to classes, it provides evidence that a good set of attributes is used to separate the classes. Note that this is a sufficient, not a necessary condition – in theory it is unlikely, but still possible, that classes are easily separable, but distributed over a multitude of clusters. However in our case there is a very good mapping from clusters to classes. It is also interesting to note that if we use two clusters on the three class problem, both classes with relatively thick cells (7936 and 6955) are grouped into a single cluster.

Finally, classifiers were built for the two-class and three-class classification problem, see Table 3 and Table 4 for the results. Classifiers were built using different combinations of attributes and classifiers; the base set of attributes is described in section 2.6. The attribute set was varied across using binary (description ends with b), grey level (ends with g) or both binary and grey value features (ends in all). Furthermore, we differentiate between using all base features, relative area only, relative thickness only and moments only. It is very clear from the two class results that image classes seem to be perfectly separable. As highlighted before, this does not guarantee that we can perfectly distinguish between the two classes of cells because in principle there could be other causes for differences between images. That said, from visual inspection it is clear that both classes of cells are quite different. Furthermore, the cluster experiments have shown that the two most similar classes end up in a single cluster when forced by the clustering algorithm (i.e. 7936, 6955). Note furthermore that a simple single split is sufficient for good performance (decision stump) and that this result is robust across the various sets of attributes. It is actually quite common in both “very hard” and “very easy” real world data mining problems that simple models produce accurate and robust results[11,20].

To make the classification a bit less trivial we have also built classifiers to separate all three classes. As can be seen from Table 4 the classification accuracy goes down for many attribute set – classification algorithm combinations. However, for some it is still possible to get near perfect results. It is interesting to note that the classifiers on moment invariants generally perform better than classifiers built on metrics that more or less directly aim to measure capsule thickness (relative area of the inner part; thickness of the capsule). This demonstrates that there is more to a cell than just the capsule thickness. Furthermore, grayscale features seem to perform better than features derived from binary images. This suggests there is more to a cell image than simple binary shape. The relatively high performance of 1 nearest neighbor is also noteworthy. We do not have an explanation for this, however, 1-nn tends to perform well if there are ‘exceptions to the rule’ that are actually not just outliers, but valid examples of a class.

4 DISCUSSION

This paper describes the development of a classification system for a capsulated pathogenic yeast. Apart from being successful in its own right it can serve as an example to other such systems in the ingredients that are used and the emphasis on each of the parts that make up the system: i.e., the specimen preparation, the image acquisition, the image processing and analysis and the classification including feature selection.

Table **3:** Classification accuracy and standard deviation (ten fold ten runs) for various combinations of attributes and classifiers (two class problem)

											with CFS attribute selection					
Description	**d. stump**		**j48/c45**		**n. bayes**		**1-nn**		**5-nn**		**n. bayes**		**1-nn**		**5-nn**	
	%acc	σ	%acc	σ	%acc	σ	%acc	σ	%acc	σ	%acc	σ	%acc	σ	%acc	σ
Base set, all	99.4	1.9	99.0	2.3	99.5	1.5	100.0	0.0	100.0	0.0	99.7	1.2	100.0	0.0	100.0	0.0
Base set, b	98.5	2.5	99.0	2.3	99.5	1.5	100.0	0.0	100.0	0.0	100.0	0.0	100.0	0.0	100.0	0.0
Eq. 4a, 4b (CB, CC),b	98.5	2.5	99.0	2.3	97.8	2.9	98.5	2.5	98.6	2.8	97.8	2.9	98.5	2.5	98.6	2.8
Eq. 6a, 6b (CC),b	98.5	2.5	99.0	2.1	98.0	2.9	98.3	2.7	98.6	2.8	98.0	2.9	98.3	2.7	98.6	2.8
Eq. 10-13, b	98.5	2.5	99.0	2.3	98.9	2.3	99.5	1.6	99.5	1.5	100.0	0.5	100.0	0.0	99.0	2.0
Base set, g	100.0	0.0	99.4	1.9	100.0	0.0	100.0	0.0	100.0	0.0	100.0	0.0	100.0	0.0	100.0	0.0
Eq. 4a, 4b (CB, CC),g	99.4	1.6	98.9	2.3	98.6	2.5	99.0	2.1	99.2	1.9	98.6	2.5	99.0	2.1	99.2	1.9
Eq. 6a, 6b (CC),g	98.5	2.5	98.9	2.5	98.0	2.9	98.5	2.5	98.5	3.0	98.0	2.9	98.5	2.5	98.5	3.0
Eq. 10-13, g	100.0	0.0	99.4	1.9	100.0	0.0	100.0	0.5	100.0	0.5	100.0	0.0	100.0	0.5	100.0	0.5
Eq. 10-13, all	99.4	1.9	99.0	2.3	99.5	1.5	100.0	0.0	100.0	0.5	100.0	0.0	100.0	0.5	100.0	0.5

Table **4:** Classification accuracy and standard deviation (ten fold ten runs) for various combinations of attributes and classifiers (three class problem)

											with CFS attribute selection					
Description	**d. stump**		**J48/c45**		**n. bayes**		**1-nn**		**5-nn**		**n. bayes**		**1-nn**		**5-nn**	
	%acc	σ	%acc	σ	%acc	σ	%acc	σ	%acc	σ	%acc	σ	%acc	σ	%acc	σ
Base set, all	76.2	1.7	95.1	3.7	91.9	4.8	96.2	3.4	94.5	4.0	93.4	4.9	97.0	3.3	96.1	3.6
Base set, b	74.3	2.9	91.7	5.0	87.6	5.6	91.8	4.5	92.6	4.7	88.2	4.9	93.0	4.3	94.1	3.8
Eq. 4a, 4b (CB, CC),b	74.3	2.9	73.8	3.6	75.0	5.6	82.1	5.8	78.3	6.6	75.0	5.6	82.1	5.8	78.3	6.6
Eq. 6a, 6b (CC),b	74.3	2.9	73.6	3.5	73.0	5.6	78.1	6.4	78.2	6.4	73.0	5.6	78.1	6.4	78.2	6.4
Eq. 10-13, b	74.5	2.8	81.7	5.9	83.1	5.9	81.7	5.8	81.3	6.1	81.6	5.9	82.0	6.3	82.9	6.3
Base set, g	76.2	1.7	94.9	3.8	94.6	4.4	97.3	3.0	95.2	3.7	96.2	3.8	97.6	2.7	96.7	3.0
Eq. 4a, 4b (CB, CC),g	72.9	3.6	84.0	5.7	79.6	6.5	79.7	6.8	83.6	6.1	79.6	6.5	79.7	6.8	83.6	6.1
Eq. 6a, 6b (CC),g	74.5	2.8	73.4	3.5	73.4	5.4	77.3	7.4	77.6	6.6	73.4	5.4	77.3	7.4	77.6	6.6
Eq. 10-13, g	76.2	1.7	88.1	4.9	87.3	5.4	83.8	6.1	81.3	5.8	87.3	5.4	83.8	6.1	81.3	5.8
Eq. 10-13, all	76.2	1.7	90.1	5.7	84.0	5.6	81.2	6.6	84.6	6.0	82.2	6.0	83.6	5.3	84.4	5.8

The outcome for the different features with respect to the classification is intriguing. If we go by the judgment of the yeast biologist we would have to let the capsule features relating to geometry dominate. This is, however, not unambiguously found in the feature selection. Rather than binary features, the first and second moment invariants are shown to be important in the classifications. One should realize that these features were derived from a masked grey-value image, consequently, for these features the density distribution is only considered at the masked geometry, i.e., CB and CC (cf. Fig. 3E). The precise mechanisms that make features discriminative can not be concluded from these experiments. New, controlled, experiments should be designed to get further insight in these features.

The three class experiments portray interesting aspects. The third strain (6955) corresponds to the *C. neoformans* stain (7936) in that it has clearly visible (thick) capsules; genetically, however, these strains are different. The imaging conditions of 7926 and 7936 were more or less similar; the 6955 strain was captured under different conditions. One could be tempted to conclude that the results for the three clusters and the classification experiments may be weakened by this fact, since the classification and clustering models could be focusing on imaging conditions rather than yeast cell differences. Yet, we observe that when the three classes are forced into two clusters, the 6955 class is correctly grouped with the 7936 class, which indicates that the models and underlying attributes are detecting cell differences, not just differences in imaging conditions. On the basis of these experiments we can not draw definitive conclusions on the influence of the imaging conditions. It would, however, certainly be interesting to further investigate which of the features are truly invariant.

In our experiments we have controlled for all other parameters than the particular yeast strain used (yeast strain production, staining, image acquisition, automated segmentation and feature analysis etc.). This allows us to conclude us with appropriate level of confidence that with our methods we are not just differentiating between images, but also between the yeast strains. That said, large historical image collections are generally not produced under controlled conditions. An important next step would be to make the classification problem more complex by adding more noise - using images that have been captured under a wide variety of conditions. This will require more robust classifiers and methodologies to ensure that true differences between classes are learned and not just 'accidental' differences between images.

The results presented in this paper are based on bright field images of negatively stained specimen preparations. Recently, antibodies against the cryptococcal capsule have become available. If these antibodies are tagged with fluorescent dyes, the acquisition can be done based of fluorescence microscopy and consequently the segmentation procedure can be further improved and simplified. Moreover, instead of the 2D approach that has been applied in our experiments, Confocal Laser Scanning Microscopy (CLSM) could be used to be able to include 3D-features. The issue of separation of cells can, in that case, be solved by other means. The use of 3D-images will, however, require using other features, as the approach taken can not be translated directly from the 2D-case.

In this paper we have focused on the image features to discriminate between potentially virulent and non-virulent cells. From the point of view of content-based multimedia retrieval we will be moving in the direction of solutions where the yeast biologist actively includes more and different data in the analysis. This requires that researchers are able to have a lot of different search, navigation and browsing dimensions to access the data. Some of these will be lower level, syntactic, but others will be more high level semantic categories (like virulent/non virulent) that have an important connotation to the biologist. Prior to the analysis, data should be submitted to a database that will incorporate direct links to the relevant bio-molecular repositories[2]. With respect to classifiers this will dictate these to be built for a wide array of classes. Ideally, automated classification procedures can be developed so that an end user, like a biologist, instead of a data miner, can create and train classifiers.

The tool presented here will allow automated analysis of capsular characteristics of many cryptococcal cells of isolates of different phenotypic or genetic background. This will be particularly useful in this OMIC-era where gene knock out strains of *C. neoformans* are being prepared and need to be analyzed for pathogenicity-related features. As stated earlier, the capsule is one of the most important characteristics to that respect. Furthermore, automated feature extraction and comparison of capsular characteristics will allow integrative studies where capsular characteristics are being compared with other features, which may be either phenotypic or genetic in nature. Among these are rates of melanization, expression profiles of virulence-related proteins, growth rates at different temperatures and substrates, assimilation patterns of carbohydrates, nitrogen compounds or vitamins, susceptibility to antifungals, genotypic data on the various subtypes known to exist in the species, and more importantly, extensive collections of transcriptome data as revealed by microarray analysis.

5 CONCLUSIONS

In this paper we have presented a holistic overview of an end to end process for classifying yeast cells using image features with the ultimate goal to detect pathogen conditions. Previous studies were based on manual measurement of capsule thickness and cell area in the binary image, but no automated procedures existed.

By carefully controlling the conditions we were able to show that through a largely automated procedure we could distinguish between a yeast strain and its mutant (7936, 7926), which simulate respectively pathogenic or non-pathogenic cells. We have shown that there are more predictive features than simply thickness and area in the binary image, some related to the density distribution in the image, or under the shape of interest, in particular the first and second moment invariants. Furthermore, we have shown that when we introduce noise in the form of a third class of a distinctly different strain this is clustered in the proper class and classifiers that need to distinguish between the three classes still achieve acceptable accuracy.

We have identified various application scenarios and associated challenges for extending the solution, such as identifying more semantically interesting classes beyond pathogenicity, making classifiers more robust for heterogeneous sets of images and developing methodologies enabling biologists rather than data miners to develop classification modules for the purpose of content based image classification, annotation and retrieval.

ACKNOWLEDGEMENTS

This work was partially supported (Ferry Hagen) by the renewal-fund of the Royal Netherlands' Academy of Sciences (KNAW) and by BioRange BSIK-research grant (Heterogeneous data analysis). In Addition, we wish to acknowledge Arco Vos and Paul Boon for their initial work on the segmentation of yeast cell images. Their work was partially supported by the STW Technology Foundation and the Royal Netherlands' Academy of Sciences.

REFERENCES

1. I. Bose, A.J. Reese, J.J. Ory, G. Janbon and T.L. Doering "A yeast under cover: the capsule of *Cryptococcus neoformans*", *Eukaryot. Cell* 2(4), 655-663 (2003).
2. Y. Bei, M. Belmamoune and F. J. Verbeek, "Ontology and image semantics in multimodal imaging: submission and retrieval", *Proc. of SPIE Internet Imaging* VII 6061, C1-C12 (2006).
3. M. Belmamoune and F. J. Verbeek, "Heterogeneous Information Systems: bridging the gap of time and space. Management and retrieval of spatio-temporal gene expression data", *Proc. Int. Conf. Multidisciplinary Information Sciences & Technologies* 1, 53-58 (2006).
4. A. Casadevall and J.R. Perfect, *Cryptococcus neoformans*, ASM Press, Washington, 1998.
5. Y.C. Chang and K.J. Kwon-Chung, "Complementation of a capsule-deficiency mutation of *Cryptococcus neoformans* restores its virulence", *Mol. Cell. Biol.* 14(7), 4912-4919 (1994).
6. M.A. Dykstra, L. Friedman and J.W. Murphy, "Capsule size of *Cryptococcus neoformans*: control and relationship to virulence", *Infect. Immun.* 16(1), 129-135 (1977).
7. W.L. Golubev and A.R. Manukyan, "Capsule formation by a saprophytic yeast", *Mikrobiologiya* 48, 314 (1979).
8. R.C. Gonzales and R.E. Woods, *Digital Image Processing*, Prentice Hall, New Jersey, 1993.
9. M.A. Hall and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining", *IEEE Transactions on knowledge and Data Engineering* 15 (6), 1437-1447 (2003).
10. M.A. Hall, "Correlation-based feature subset selection for machine learning", University of Waikato, Hamilton, New Zealand (1999)
11. R. Holte, "Very simple classification rules perform well on most commonly used datasets", *Machine Learning* 11, 63–91 (1993).
12. M. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Inf. Theory* 8, 179-187 (1962).
13. G. Janbon, "*Cryptococcus neoformans* capsule biosynthesis and regulation", *FEMS Yeast Res.* 4(8), 765-771 (2004).
14. R. Kohavi, "The power of decision tables", *Proc European Conference on Machine Learning* (1995).
15. M.L. Littman, "Capsule synthesis by *Cryptococcus neoformans*", *Trans NY Acad Sci.* 20(7), 623-648 (1958).
16. M.L. Littman and E. Tsubura, "Effect of degree of encapsulation upon virulence of *Cryptococcus neoformans*", *Proc. Soc. Exp. Biol. Med.* 101, 773-777 (1959).
17. J. Liu, P. van der Putten, F. Hagen. X. Chen, T. Boekhout and F.J. Verbeek, "Detecting virulent cells of *Cryptococcus neoformans* yeast: clustering experiments", *Proceedings ICPR* Volume 1 IEEE Computer Society 2006, 1112- 1115, Hongkong, China (2006).
18. J. Rivera, M. Feldmesser, M. Cammer and A. Casadevall, "Organ-dependent variation of capsule thickness in *Cryptococcus neoformans* during experimental murine infection", *Infect. Immun.* 66(10), 5027-5030 (1998).

19. R. van Balen, D. Koelma, T.K. ten Kate, B. Mosterd and A.W.M. Smeulders, "ScilImage: A multi-layered environment for use and development of image processing software", in: *Experimental environments for computer vision and image processing*, H.I.C.J.L. Christensen, ed., World Scientific, Singapore (1993).
20. P. van der Putten and M. van Someren. "A Bias-Variance Analysis of a Real World Learning Problem: The CoIL Challenge 2000". Machine Learning, vol. 57, iss. 1-2, 177-195, Kluwer Academic Publishers (2004)
21. F.J. Verbeek, *Three Dimensional reconstruction of biological objects from serial sections including deformation correction*, Delft University of Technology, Delft, The Netherlands, (1995).
22. F.J. Verbeek, "Theory & Practice of 3D-reconstructions from serial sections", in: *Image Processing, A Practical Approach.* R.A. Baldock and J. Graham, eds. Oxford: Oxford University Press (1999).
23. I.H. Witten and E. Frank, *Data mining: practical machine learning tools and techniques with Java implementations*, Morgan Kaufmann Publishers, San Francisco, 1999.
24. O. Zaragoza, B.C. Fries and A. Casadevall, "Induction of capsule growth in *Cryptococcus neoformans* by mammalian serum and CO(2)", *Infect. Immun*. 71(11), 6155-6164 (2003).
25. O. Zaragoza and A. Casadevall, "Experimental modulation of capsule size in *Cryptococcus neoformans* Biological", *Procedures Online* 6, 10-15 (2004).

[#] In the text we have abbreviated the yeast strains to just numbers; thus 7936 is identical to CBS 7936, 7926 is identical to CBS 7926 and 6955 is identical to CBS 6955.