ELSEVIER

Journal of Microbiological Methods 52 (2003) 221–229

Journal of Microbiological Methods

www.elsevier.com/locate/jmicmeth

Visual clone identification of *Penicillium commune* isolates

Michael Edberg Hansen [a,*], Flemming Lund [b], Jens Michael Carstensen [a]

[a] *Informatics and Mathematical Modelling, Richard Petersens Plads, Building 321, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark*

[b] *BioCentrum-DTU, Søltofts Plads, Building 221, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark*

Received 4 January 2002; received in revised form 10 July 2002; accepted 29 July 2002

Abstract

A method for visual clone identification of *Penicillium commune* isolates was developed. The method is based on images of fungal colonies acquired after growth on a standard medium and involves a high degree of objectivity, which in future studies will make it possible for non-experts to perform a qualified identification of different species as well as clones within a species. A total of 77 *P. commune* isolates from a cheese dairy were 3-point inoculated on Yeast Extract Sucrose (YES) agar and incubated for 7 days at 25 °C. After incubation, the isolates were classified into groups containing the same genotype determined by DNA fingerprinting (AFLP). Each genotype also has a specific phenotype such as different colony colours. By careful image acquisition, colours were measured in a reproducible way. Prior to image analysis, each image was corrected with respect to colour, geometry and self-illumination, thereby gaining a set of directly comparable images. A method for automatic extraction of a given number of concentric regions was used. Using the positions of the regions, a number of relevant features—capturing colour and colour–texture from the surface of the fungal colonies—was extracted for further analysis. We introduced the Jeffreys–Matusitas (JM) distance between the feature distributions to express the similarity between regions in two colonies, and to evaluate the overall (weighted) similarity. The nearest neighbour (NN) classification rule was used. On a dataset from 137 isolates, we obtained a "leave-one-out" cross-validation identification rate of approximately 93–98% compared with the result of DNA fingerprinting.

Keywords: Direct identification; *Penicillium commune*; Filamentous fungi; Image analysis; Macromorphology; Cross-validation

1. Introduction

Penicillium commune is the most widespread and most frequently occurring spoilage fungi on cheese (Lund et al., 1995). Growth of *P. commune* on cheese results in discolouring of the surface and production of off flavors. To reduce contamination, cheese dairy staff disinfect the production environments daily; however, unwanted growth of *P. commune* continues to be a problem. In a study by Lund et al. (submitted for publication) to map contamination routes, 321 *P. commune* isolates were sampled from cheese and the production environment of two dairies, and characterized by phenotype and DNA fingerprinting techniques like RAPD and AFLP. *P. commune* isolates were classified into groups according to both pheno- and AFLP genotype. Genotypic similarity indicated that the isolates were clones and were assumed to have a common origin in the dairy. In this way, a clone in the dairy could be traced and the contamination points determined. When working with these *P. commune*

* Corresponding author.
E-mail address: meh@imm.dtu.dk (M.E. Hansen).

0167-7012/03/$ - see front matter © 2003 Elsevier Science B.V. All rights reserved.
PII: S0167-7012(02)00160-4

isolates, it was noticed that different isolates on standard identification media had different appearances, i.e. colonies could have different diameter or colour. Lund et al. (submitted for publication) made a comparative study of the AFLP data and different phenotypes found as subjective visual evaluation of different colony morphology and colours. The results of the experiment were reproducible even when non-experts made a visual examination of the colonies. A total of 272 isolates were investigated and a comparison between the phenotypes and the AFLP method showed a correct identification rate of 95%. In spite of these good results though, it is always better to use objective methods.

Dörge et al. (2000) have shown that it was possible to identify *Penicillium* species using image analysis. Careful image acquisition under standardized conditions will provide very accurate and reproducible colour images. Processing this data is possible by storing visual formulations in a phenotype database, creating an opportunity for explorative data analysis to reveal new information.

In this work, we compared results from DNA fingerprinting with results from image analysis. The objective was to investigate if image analysis could support or maybe serve as a substitute for subjective phenotyping methods and to substantiate the DNA fingerprinting of *P. commune* isolates.

2. Materials and methods

The fungal isolates was obtained from the IBT culture collection, BioCentrum-DTU, Technical University of Denmark.

A total of 77 *P. commune* isolates from one cheese dairy were investigated. All isolates were 3-point inoculated on Yeast Extract Sucrose (YES) agar using Difco Yeast extract (for formulae, see Samson et al., 2000) and incubated in the dark for 7 days at 25 °C in ventilated plastic bags. Colonies were inoculated the same day on the same batches of media.

Table 1 shows how the fungi were grouped into 16 classes by DNA fingerprinting. The X class contains 27 individual isolates that could not be paired with any other. A full list with low resolution copies of the images, is available from the World Wide Web: www.imm.dtu.dk/~meh.

3. Digitization

In the digitization procedure, colours as they appear are mapped into discrete arrays of pixel values representing a digital image, *I*. One of the first observations to make was that even though the visual variation across the whole ensemble of isolates was

Table 1
Classification of 77 *P. commune* isolates into 15 classes (A–O) containing identical clones according to DNA fingerprinting

Class	IBT numbers	Number of isolates
A	5 × IBT 23588 IBT 23589 IBT 23656 IBT 23657 IBT 23658 IBT 23659 IBT 23660 IBT 23661 IBT 23662 IBT 23663 IBT 23664	15
B	5 × IBT 23586 IBT 23587 IBT 23595 IBT 23596	8
C	5 × IBT 23592 IBT 23613	6
D	5 × IBT 23624 IBT 23625 IBT 23637 IBT 23649	8
E	5 × IBT 23605 IBT 23606 IBT 23612	7
F	5 × IBT 23627 IBT 23647	6
G	5 × IBT 23635 IBT 23648	6
H	5 × IBT 23650 IBT 23653 IBT 23629 IBT 23628	8
I	5 × IBT 23590 IBT 23591	6
J	5 × IBT 23645 IBT 23646 IBT 23655	7
K	5 × IBT 23598 IBT 23622	6
L	5 × IBT 23608 IBT 23652 IBT 23599 IBT 23609 IBT 23610	9
M	5 × IBT 23602 IBT 23616	6
N	5 × IBT 23543 IBT 23579	6
O	5 × IBT 23640 IBT 23644	6
X	IBT 23651 IBT 23634 IBT 23638 IBT 23633 IBT 23623 IBT 23621 IBT 23620 IBT 23607 IBT 23614 IBT 23603 IBT 23642 IBT 23593 IBT 23594 IBT 23600 IBT 23604 IBT 23611 IBT 23615 IBT 23617 IBT 23618 IBT 23619 IBT 23626 IBT 23631 IBT 23632 IBT 23636 IBT 23639 IBT 23641 IBT 23654	27
		Total 137

Some of the isolates were replicates and the total number of tested isolates on Petri dishes was 137. The X class contains 27 individuals.

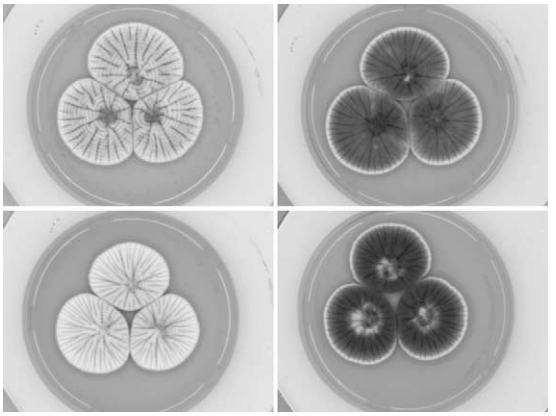

Fig. 1. Example of four Petri dishes after digitization. All isolates are different clones of *P. commune* and illustrate some of the variation across the whole ensemble of isolates.

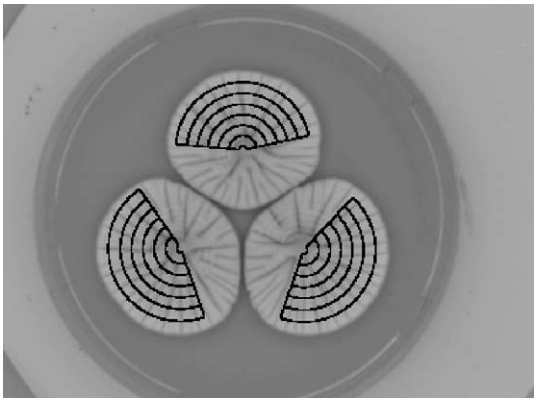

Fig. 2. Output after pre-processing. The colonies have been identified and subregions extracted. From each of the regions, features are extracted and used for analysis.

large—both with respect to morphology and colour, some groups were quite similar. This means that the visual difference between two genetically different groups could be insignificant. To measure these subtle differences in colour and texture, it was of crucial importance that we controlled all parameters affecting the outcome of the digitization. We used the same digitization method as proposed by Dörge et al. (2000) (Fig. 1).

4. Region extraction

The obtained images contained a large amount of information. To reduce the computational complexity, we first detected the Petri dish, then the colonies and inoculation points. These areas will be referred to as Regions of Interest (ROIs) (Sonka et al., 1998).

In Dörge et al. (2000), a method for the automatic segmentation of both the Petri dish and colonies was proposed and tested on a variety of images. An improved version of this method has been implemented and used to segment the images into Petri dish and further into a number of regions containing each colony. The algorithm was fully automated, ensuring that the extraction was done consistently for all images. After segmentation, each colony-containing region was divided further into K concentric regions. If $K=1$ the region contained the whole colony and if $K=3$ the colony was divided into three regions. In Fig. 2, results after ROI detection are shown; here, $K=6$ has been chosen for illustrative purposes.

5. Feature extraction

The method proposed in Dörge et al. (2000) for direct identification of fungal species was tested. The method turned out to have low performance on the clones in this study due to the high degree of visual similarity, and new features were needed. Our intent was to evaluate the direct visual information available in the colonies, therefore requiring a new set of features for analysis.

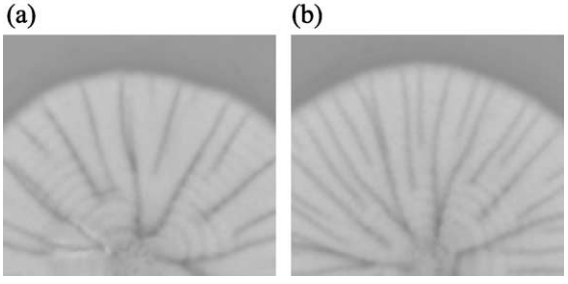

Fig. 3. Figure illustrating the variety and some of the features that are going to be used in the identification procedure. The two isolates have approximately the same mean RGB colour value, but their rippledness is significantly different.

5.1. Colour

The first feature used was based on calibrated (RGB) colour measurements extracted from each of the pixels in the K regions inside the colonies. The features from the same regions in different colonies, inside a Petri dish, were pooled to account for natural variations.

5.2. Texture—topology/morphology

Even though colours (intensities) constitute the most important factor for the human visual system of identification (Landy, 2002), spatial distribution is also important for the perception and understanding of a scene. Different isolates may have the same global content of pixels having certain colours, but the spatial relation of the pixels determines how we identify them.

In Fig. 3, two very similar colonies are shown. We see that although some isolates look somehow similar and contain the same overall distribution of colour values, we are still capable of distinguishing each of them. What differs is how pixels are organized (how they are distributed), which results in a different pattern (texture). What is common for all isolates is that the patterns are more or less radial or angular with the center at the inoculation point. By using this information about symmetry, we have a good feature discrimination on the level of rippledness in these two directions. Many methods exist for capturing texture. Since we want to describe the density of colour in a nearby region, one simple method could be to evaluate the gradient in the proper directions (Fig. 4(a)).

In mathematical terms, the expression for the gradient (in cartesian/rectangular coordinates) at a given point $\boldsymbol{x}=(x_1, x_2)$ in the image, may be calculated using the *gradient operator*, ∇:

$$\nabla I(\boldsymbol{x}) = \begin{bmatrix} \frac{\delta I}{\delta x_1} \\ \\ \frac{\delta I}{\delta x_2} \end{bmatrix} \left(= \frac{\delta I}{\delta x_1}\hat{x}_1 + \frac{\delta I}{\delta x_2}\hat{x}_2 \right) \qquad (1)$$

Here, I is a generalization of the image, and hence $I(x_1, x_2)$ is the pixel value at the coordinate (x_1, x_2) in the image. The gradient estimates the change in the colour intensity values when going in certain directions. The gradient operator in Eq. (1) estimates the gradient in both the horizontal and vertical directions. Unfortunately, the structures we are interested in are concentric. To capture the rotational and radial changes, we need to transform our gradients. At any point inside the colony, the direction and angle to the inoculation center are given by an angle, θ, and direction, ρ (Fig. 4(a)). Based on θ, it is possible to transform (a change of basis) the horizontal/vertical gradients to circular symmetric gradients. At a point from distance ρ in the direction θ from the inoculation

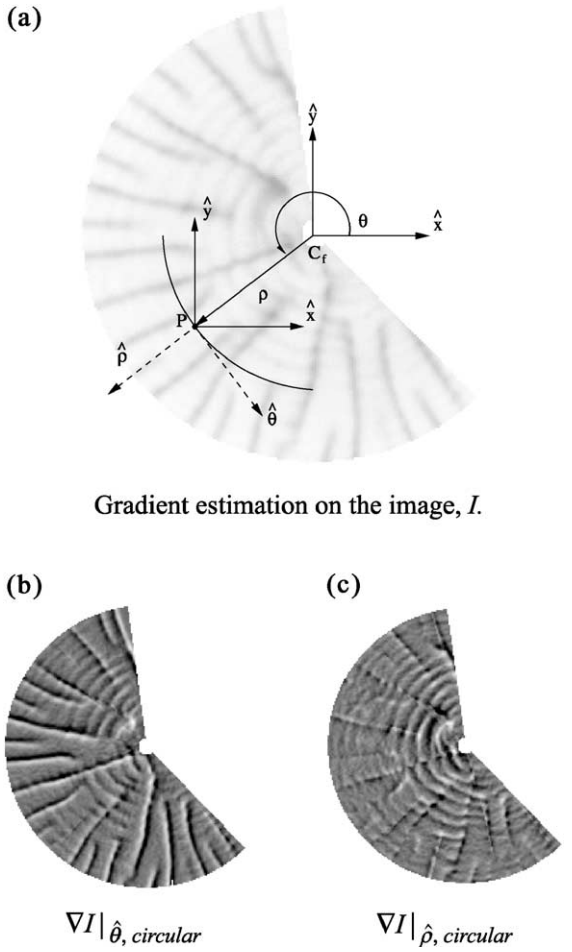

Fig. 4. This figure illustrates the principle of the gradient based feature enhancement. In panel (a), $(\hat{x}_1, \hat{x}_2)$ and $(\hat{\rho}, \hat{\theta})$ span out the rectangular and circular coordinate systems. C_f is the inoculation point for the colony. ρ and θ represent the distance and angle, respectively, from the center to a given point, P, within the colony. From Eq. (2), the gradients are estimated. The resulting gradients in both directions are illustrated in panels (b) and (c).

point, the radial and angular gradients are estimated by

$$\nabla I(\boldsymbol{x})\,|_{\text{circular}} = \boldsymbol{\Gamma}\cdot\nabla I(\boldsymbol{x})\,|_{\text{cartesian}} \tag{2}$$

where $\boldsymbol{\Gamma}$ is the rotation matrix

$$\boldsymbol{\Gamma} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{3}$$

In Fig. 4, the result of the estimated gradient, $\nabla I(\boldsymbol{x})\,|_{\text{circular}}$ in both the angular (Fig. 4(b)) and radial (Fig. 4(c)), directions are shown.

We found that the ripples in the colony surface resulted in high gradient values and that the they were well separated into radial and angular gradients. After the feature extraction, we now had a set of parameters describing some properties of the colonies. Based on these features, the strategy was to evaluate the degree of similarity between the feature distributions.

6. Similarity

A similarity function is a function that assigns a nonnegative real number to each pair of clones defining a notion of resemblance. Similarity measures form the basis for many matching algorithms, and a more detailed introduction to the subject can be found in Gordon (1999) and Androutsos et al. (1998).

Fungal isolates can be described by a set of features where the features span out dimensions. An isolate is then represented in the feature space as a cloud of points extracted from the colony. The similarity then has to be evaluated as a distance between the distributions of features, and the task is to define a proper metric in which we may calculate differences.

If we had isolates that could be sorted on the basis of the term "dark" and "bright", the mean value of the feature distributions would be sufficient. The real world has made the task more difficult, and some clones may have the same mean colour (red, green and blue) value.

6.1. Similarity measure

The Jeffreys–Matusitas distance (JM distance) (Matusita, 1956) has been shown to provide a reliable criterion as a function of class separability or similarity. For two isolates having features with probability densities $p_i(\boldsymbol{x})$ and $p_j(\boldsymbol{x})$, the JM distance, J_{ij}, is given by

$$J_{ij} = \left[\int_{\boldsymbol{x}} \left(\sqrt{p_i(\boldsymbol{x})} - \sqrt{p_i(\boldsymbol{x})}\right)^2 \mathrm{d}\boldsymbol{x}\right]^{1/2} \tag{4}$$

The most common method is a parametric approach in which observations are assumed to belong to a certain type of distribution. However, sometimes the densities are of a much more complex nature, and hence it is not possible to have a simple parametric model for the densities. One way of overcoming this is by using non-parametric estimates of the densities.

Let us assume that the features we have extracted from two isolates, i and j, may be approximated by two normal distributions with the mean values $\hat{\boldsymbol{\mu}}_i$, $\hat{\boldsymbol{\mu}}_j$ and the dispersions matrices $\hat{\boldsymbol{\Sigma}}_i$, $\hat{\boldsymbol{\Sigma}}_j$ of the features. In this case, the JM distance can be simplified and expressed by the parameters in the model, and Eq. (4) reduces to

$$J_{ij} = \sqrt{2(1-\mathrm{e}^{-\alpha_{ij}})} \qquad J_{ij} \in \left[0;\sqrt{2}\right] \tag{5}$$

where α_{ij} is called the Bhattacharyya distance and is covered in many texts on statistical pattern recognition (Fukanaga, 1990). α_{ij} describes the standardized distance between isolates i and j and based on the estimated means and dispersions we calculate the Bhattacharyya distance, α_{ij}, as

$$\alpha_{ij}^2 = \underbrace{\frac{1}{8}\,(\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_i)^t \hat{\Sigma}^{-1} (\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_i)}_{A} + \underbrace{\frac{1}{2}\ln\frac{|\hat{\Sigma}|}{\sqrt{|\hat{\Sigma}_i|\,|\hat{\Sigma}_j|}}}_{B} \tag{6}$$

where

$$\hat{\Sigma} = \frac{\hat{\Sigma}_i + \hat{\Sigma}_j}{2} \tag{7}$$

is the average dispersion for both the distributions. In Eq. (6), the first term, A, corresponds to the distance between the mean values assuming the distributions have the same dispersion. In the literature, this dis-

tance is often referred to as the *Mahalanobis* distance. The latter term, B, expresses the difference in the dispersion matrices.

An illustration of the JM distance, can be seen in Fig. 5. Here, the curves plot the values of two density functions (artificial functions for illustration purpose), $\sqrt{p_i(\boldsymbol{x})}$ and $\sqrt{p_j(\boldsymbol{x})}$, and their joint contribution to the JM distance (gray area). Fig. 5(a) and (b) shows two special cases where the first illustrates the case where the distributions have different means, and the latter where they have similar mean. Dispersion is different in both cases. The JM distance is then illustrated as the gray area. In the latter case, where the mean vector is similar for each of the feature sets above, the Mahalanobis distance will be zero due to the terms of difference between the mean values. As long as the observations are spread differently, the second term contributes to α_{ij}. Only if mean values and dispersion matrices are exactly identical do we have $\alpha_{ij}=0$.

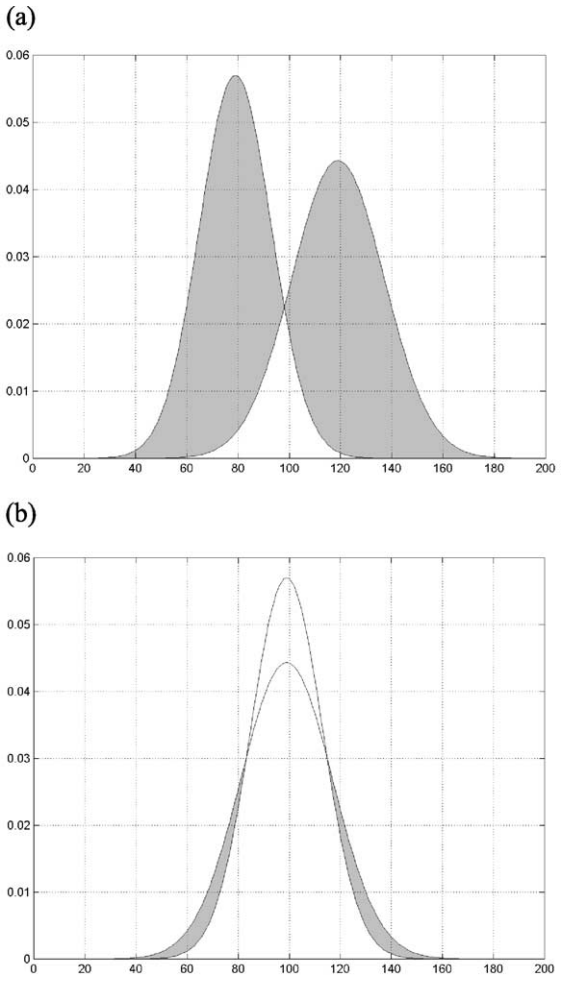

Fig. 5. Illustration of the joint contribution to the JM distance (Eq. (5)) between two distributions. In both cases, we have $\hat{\boldsymbol{\Sigma}}_i \neq \hat{\boldsymbol{\Sigma}}_j$. It is important to note that even when the mean value of the distribution is similar, the different dispersions also contribute. (a) Distribution with different means and dispersions, $\hat{\boldsymbol{\mu}}_i \neq \hat{\boldsymbol{\mu}}_j$. (b) Distribution with equal means and different dispersions, $\hat{\boldsymbol{\mu}}_i \neq \hat{\boldsymbol{\mu}}_j$.

6.2. Region weighting

Using the JM distance in this context, we get a vector of distances

$$\boldsymbol{d}_{ij} = \boldsymbol{J}_{ij} \tag{8}$$

where $\boldsymbol{d}_{ij}=(d_{ij,1},\ldots,d_{ij,K})^t$, is the vector of distances calculated on features extracted from the K regions, between two isolates, i and j.

A natural choice for combining the elements of the vector into one overall similarity, is to create a linear combination

$$\begin{aligned} d_{\text{overall}} &= g(\boldsymbol{d};\ \boldsymbol{w}) \\ &= \boldsymbol{w}^t\boldsymbol{d} = \sum_k w_k d_k \end{aligned} \tag{9}$$

where $\boldsymbol{d}$ is the vector of regional distances and $0 \leq w_k \leq 1$ are weights.

7. Data analysis

Having done the localization of each of the colonies in all images, we obtained a data set of observations with a distribution within *P. commune* as shown in Table 1. For each of the observations, the features described in Section 5 were measured and stored. Based on these features, it is possible to evaluate the degree of similarity between isolates based on the method described in Section 6.

7.1. Identification

In the identification procedure, the task was to compare an unidentified isolate with a set of known

classified isolates. For references, the region and feature extraction could be done once and for all, and so the only steps left would be to do the extraction on the unidentified isolate, and then calculate the distances to each of the known isolates, based on the JM distance. The estimated distances are combined into one overall similarity, according to Section 6.2. In this study, we use $w_k = 1/K$. Having a set of distances between features extracted from isolates, the task is how to make a natural selection. Several methods exist, but one proper choice would be to use the *k nearest neighbour* (NN) method (Bishop, 1995). The *k*-NN identification rule is a nonparametric supervised pattern classification technique. Given a set of known isolates, the correct identification of an isolate would be to choose the class that is most heavily represented among its *k* nearest neighbours.

7.2. Cross-validation

We used the "leave-one-out" cross-validation to test the performance of the classifier. Each time we compared two isolates, we had the prior knowledge of whether or not they were genetically identical. One way to illustrate how well the distances, J_{ij}, between identical and non-identical isolates perform as classifier, is by looking at the distance distributions. Having a good classifier, we would expect the histograms to be separated. The distance to identical isolates has to be as small as possible, whereas distance to different isolates should be as large as possible. In Fig. 6, the histograms for the distances of identical and different isolates are plotted. All values lie within the interval of $[0; \sqrt{2}\,]$ (Eq. (5)).

Table 2
The identification rates for the textural, colour and combined features extracted from different number of regions

Regions, *K*	Identification rate		
	Colour (%)	Texture (%)	Combined (%)
1	93.64	92.73	95.45
2	95.45	92.73	96.36
3	96.36	93.64	97.27
6	98.18	92.73	98.18

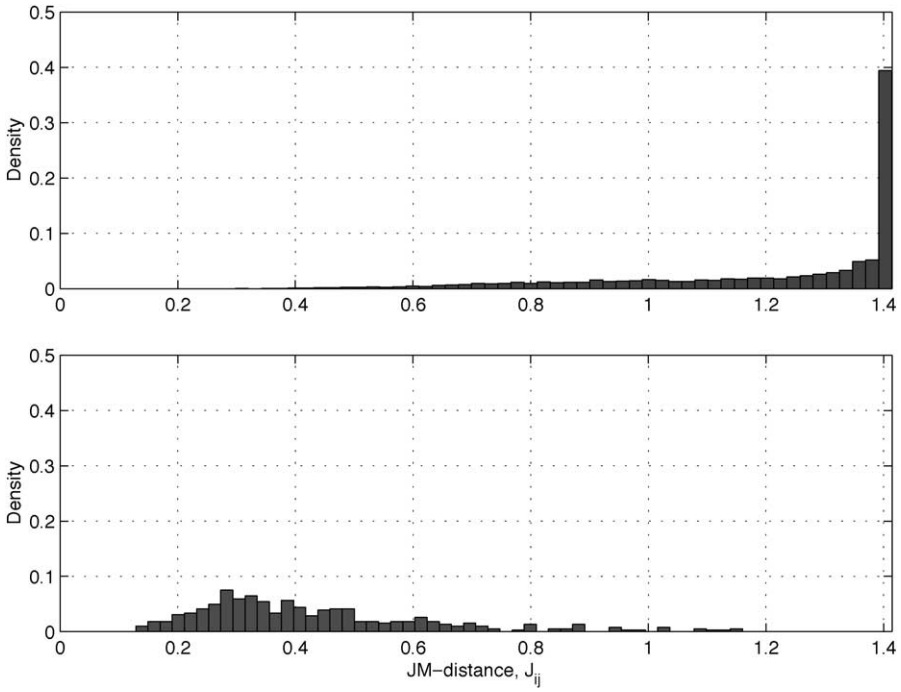

Fig. 6. Histogram of distances between the feature distributions of *different* (upper) and *identical* (lower) isolates. Having a good classifier, we would expect the histograms to be separated. The distance to identical isolates needs to be as small as possible, while the distance to different isolates needs to be as large as possible.

The final result is summarized in Table 2, and show the *identification rate* (IR) when using a different number of regions $K = 1$, 2 and 3.

We see an increase in the identification rate when comparing isolates based on features extracted from one region to two regions, and an even better result when using three or six. A further increase in the number of regions did not show a further improvement. The identification rate was high for both the colour and gradient based features.

8. Discussion

As indicated in Fig. 1, some *P. commune* isolates were very different and easy to separate. However, as isolates become more similar, visual determinations become more subjective. We have shown, that it was possible to measure objectively if two *P. commune* isolates are identical or different with a significant probability of 95–98% (see Table 2). These results were in full accordance with those obtained by Lund et al. (submitted for publication), and proves that this method could be used to confirm or even substitute for DNA fingerprinting, which in future studies will make it possible for non-experts to perform a qualified identification not only of different species but also of clones within a species.

The reasons for the success and high identification rate for both colour- and gradient-based features were mainly due to the combination of the distribution-based similarity measure, and the nearest neighbour classification rule. The similarity estimation based on a weighted distance metric and the following identification was built upon the following hypothesis: feature distributions can be expressed by a parametric expression. In this case, the model assumes that features are *normally* distributed. Although this is a rough approximation of reality, it has been proven to perform well on such isolates. One improvement would be to calculate the JM distance based on a nonparametric approach (Fukanaga, 1990). The features describing the spatial relations between the RGB pixel values were based on simple gradients, and more complex features can be applied. The weights (Eq. (9)) could also be estimated by optimizing the identification rate.

The method has been tested on a further set of data produced from the exact same isolates, but each strain was only tested once. Here, the performance of the method exceeded 93% (three regions, colour + texture) even though the number of samples within each group were smaller. Performance of the nearest neighbour classification rule improves with an increasing number of observations. This explains why we observed a slightly lower identification rate in the additional study than the first study. The results indicate that the method proposed in this paper is highly reproducible.

The isolates described in this study are restricted to *P. commune*, but the methods described above could be transferred to other *Penicillium* species or to other species of genera of fungi. This gives rise to the need for more specific features, which is a topic for future work. Furthermore, the methods developed in this study could be used for extracting isolates from a fungal database combining other types of data, e.g. visual and chemical data.

Acknowledgements

This project was supported partly by the Danish Technical Research Council under the project "Programme for predictive biotechnology: Functional biodiversity in *Penicillium* and *Aspergillus*" (grant no. 9901295), Danish Dairy Research Foundation and the Danish Government. The authors would like to thank Jens C. Frisvad, BioCentrum-DTU, and David Overy, BioCentrum-DTU, for proofreading this manuscript.

References

Androutsos, D., Plataniotis, K.N., Venetsanopoulos, A.N., 1998. Distance measures for color image retrieval. International Conference on Image Processing 2, 770–774.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Clarendon Press, Oxford.

Dörge, T., Carstensen, J.M., Frisvad, J.C., 2000. Direct identification of pure *Penicillium* species using image analysis. Journal of Microbiological Methods 41, 121–133.

Fukanaga, K., 1990. Introduction to Pattern Recognition, 2nd ed. Academic Press, London, UK.

Gordon, A.D., 1999. Classification, 2nd ed. Chapman & Hall, London, UK.

Landy, M.S., 2002. Visual perception of texture. Under review for publication in Chalupa (http://www.cns.nyu.edu/msl/texentry2.pdf).

Lund, F., Filtenborg, O., Frisvad, J.C., 1995. Associated mycoflora on cheese. Food Microbiology 12, 173–180.

Lund, F., Nielsen, A.B., Skouboe, P., 2002. Distribution of *Penicillium commune* contaminants in cheese dairies mapped using secondary metabolite profiles, phenotypes, RAPD and AFLP fingerprinting. Food Microbiology (submitted).

Matusita, K., 1956. Decision rule, based on the distance, for the classification problem. Annals of the Institute of Statistical Mathematics 8, 67–77.

Samson, R.A., Hoekstra, E.S., Frisvad, J.C., Filtenborg, O., 2000. Introduction to Food and Airborne Fungi, 6th ed. Centraalbureau voor Schimmelcultures, Utrecht.

Sonka, M., Hlavac, V., Boyle, R., 1998. Image Processing, Analysis and Machine Vision, 2nd ed. Chapman & Hall, California, USA.