Title: Extended multiplicative signal correction for FTIR spectral quality test and pre-processing of infrared imaging data

Valeria Tafintseva[1,*], *Volha Shapaval*[1], *Margarita Smirnova*[1,2], *Achim Kohler*[1]

*Corresponding Author: E-mail: valeria.tafintseva@nmbu.no

[1]Faculty of Science and Technology, Norwegian University of Life Sciences, 1432 Ås, Norway
[2] Faculty of Biology, Belarusian State University, Postbox 220030, Nezavisimosti Ave., 4, Minsk, Belarus

Keywords: spectral quality test, infrared spectroscopy, Extended Multiplicative Signal Correction

Spectral quality control is an important step in the analysis of infrared spectral data, however, often neglected in scientific literature. A frequently used quality test that was originally developed for infrared spectra of bacteria is provided by OPUS software from Bruker Optik GmbH. In this study the OPUS quality test is applied to a large number of spectra of bacteria, yeasts and moulds and hyperspectral images of microorganisms. It is shown that the use of strict thresholds for parameters of the OPUS quality test leads to discarding too many spectra. A strategy for optimizing parameters thresholds of the OPUS quality test is provided and a novel approach for spectral quality testing based on Extended Multiplicative Signal Correction (EMSC) is suggested. For all the data sets considered in our study, the EMSC quality test is shown to be the best among different alternatives of OPUS quality test provided.

1. Introduction

Fourier-transform infrared (FTIR) spectroscopy has been successfully used for characterisation and classification of microorganisms for several decades. [1-7] FTIR spectroscopy is nowadays used on different scales allowing a high-throughput characterization of different microorganisms and other cell types and exploring biochemical composition of cells at a high spatial resolution by a broad range of infrared microscopic imaging techniques. [8, 9] An FTIR spectrum represents a high-dimensional and reproducible molecular "fingerprint" of the chemical composition of biological cells and tissues containing characteristic signals from cell lipids, proteins, nucleic acids and carbohydrates. [10, 11] FTIR

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/jbio.201960112

spectral fingerprints combined with multivariate spectral analysis is therefore used frequently for the characterisation and identification of microbial cells and tissues. [[6, 12-18]

The multivariate data analysis of infrared spectra of cells and tissues consist typically of several steps: (i) a quality test of the spectra, (ii) pre-processing of the spectra to normalize and remove scatter and other unwanted instrumental effects [19-25], (iii) model establishment for, e.g., clustering, calibration or classification. [6, 15, 26] Little attention is often paid to the first step, namely the quality test of the data. The spectral quality test (QT) approaches and results are often reported very briefly or even neglected in the scientific studies. This may be due to the fact that there are no standardized routines for a quality test of FTIR spectra. The approaches to quality test FTIR spectral data and FTIR imaging data could be separated into three main groups, based on the following parameters: 1) signal-to-noise ratio; 2) peak intensity; 3) noise. [27-31] Spectra with low signal-to-noise and high noise are usually discarded. QT based on a peak intensity could be split into two different approaches: i) identifying spectra with low intensity of a relevant peak, for example amide I peak, and ii) identifying spectra with high intensity of an unwanted or irrelevant peak(s) such as water vapour in samples. Those spectra are discarded by the QT. QT of FTIR imaging data has been extended to a couple more parameters related to FTIR imaging, namely 1) “test for an additional band” corresponding to tissue embedding medium and 2) “bad pixel” test to remove spectra corresponding to dead pixels of a detector. [31] However, the choice of the selected parameters (for example, signal-to-noise, peak intensity, or noise) and their thresholds have to be defined for every data set at hand separately. Thus, there is no standard QT which can be used for a particular type of microorganism or tissue.

OPUS software provided by Bruker Optik GmbH [32] is the only standard QT and the most commonly used software for FTIR spectral quality testing. [11, 26, 33-35] When the OPUS QT is applied to infrared spectral data, parameters are calculated for each spectrum. If one of the parameters exceeds the threshold set up by the QT, the spectrum is identified as of poor quality and discarded. The most important parameters of the OPUS QT are *Absorbance (Abs)*, *Noise*, *Signal-to-Noise (S/N) ratio* and *Signal-to-*

Water (S/W) ratio. *Abs* is calculated in a range 2100 – 1600 cm^{-1}, *Noise* is calculated in a range 2000 – 2100 cm^{-1} which is chemically inactive. Two absorbance signals are used for the QT: the so-called S_1, which is calculated in a range 1700 – 1600 cm^{-1}, and S_2, which is calculated in a range 960 – 1200 cm^{-1}. The first signal represents the amide I peak at 1650 cm^{-1} referring to proteins, and the second represents polysaccharide ring vibrations at 1080 cm^{-1}. The choice of the ranges for the OPUS QT parameters are justified by the representative peaks of an infrared spectrum of microbial cells. However, a great importance is given to the amide I peak since the *Abs* parameter is calculated using exclusively this area of the spectrum (2100 – 1600 cm^{-1}). The importance given to the amide I band derives from the fact that the OPUS QT was originally established for infrared spectra of bacterial cells exhibiting strong amide I bands. However, infrared spectra of other microbial cells such as yeasts and filamentous fungi display quite different spectral bands. While the amide I and II bands from proteins dominate the infrared spectra of bacterial cells, carbohydrates and lipids are present in high concentrations in yeasts and fungi. For example, spectra from oleaginous fungi reveal bands of lipids that are much stronger than bands from proteins. [6, 14] Fungi (filamentous fungi and yeasts) have in general different morphologies and therefore different spectral characteristics influencing the quality of the spectra. Therefore, different QT strategies and parameter thresholds may be needed for a QT of infrared spectra of such cells. The same holds for imaging data: if the same threshold values that were developed for the analysis of infrared spectra of populations of bacterial data were applied for example for microspectroscopic imaging data, the major part or even all of the spectra would not pass the QT.

In this study we consider the OPUS QT for spectra of three types of microorganisms, namely filamentous fungi [11, 12], yeasts [3] and bacteria. [18] We investigate the OPUS quality parameters for a large number of spectra and compare the distributions of the parameters for filamentous fungi, yeasts and bacteria. We investigate how the quality parameters influence the classification modelling and suggest strategies for optimizing the thresholds of the parameters for a given type of microbial cell. To assess the performance of the QT two classification analysis methods were used: Partial Least Squares Regression

[36] and Random Forest. [37] Random Forest has during recent years proven to be one of the most powerful methods for classification based on infrared spectra of cells and tissues. [12, 38-40] Success rate (accuracy) of classification was used to select the best QT approach.

Further, we introduce an additional approach for quality testing of spectral data based on Extended Multiplicative Scatter Correction (EMSC). [19, 41] EMSC is a model-based pre-processing tool that allows estimating parameters related to physical, chemical and instrumental effects. [23, 42, 43] The EMSC parameters can be used to characterize morphological and chemical properties of samples. [42, 44] Among other parameters, the EMSC model estimates a parameter, commonly referred to as *b*, that refers to the effective optical thickness and morphology of samples, which can provide important physical sample information as previously shown. [42, 45] Since the parameter *b* of the EMSC model captures important information about sample morphology, we investigate further if this parameter can be used for establishing binary segmentation of FTIR images or masks. We show that background information can easily be separated from the foreground or sample regions in infrared images using such masks.

2. Materials and Methods

2.1. Data

2.1.1. FTIR spectra of filamentous fungi

A set of 59 well-characterized filamentous fungi strains from 10 different genera and 19 species obtained from the mycological strain collection of the Norwegian Veterinary Institute (Oslo, Norway) were previously measured and analyzed by FTIR spectroscopy in a study performed by Shapaval et al. [11] The filamentous fungi set was organized in a taxonomic tree with five levels of phylogenetic hierarchy (phylum, class, family, genus, and species). It is important to note that the consensus of the taxonomic tree has changed since it was published by Tafintseva et al. [12] The updated taxonomic tree is provided in Supplementary materials (**Figure S1**). The classification problem considered in this paper is the classification into groups of species using FTIR spectra.

The experimental design, used for preparing fungal spectra, was the following: growth experiments were performed in six runs - six independent experiments each performed on a separate day and for each run, all strains were cultivated twice resulting in 12 biological replicates. For each biological replicate two technical replicates were measured by the high-throughput screening eXTension unit coupled to a Tensor 27 FTIR spectrometer (both Bruker Optik GmbH, Germany).

The complete data set (1029 FTIR spectra) was used to evaluate the quality of the data, to obtain ranges of OPUS QT parameters, and to find optimal parameter thresholds. For building classifiers on species level, five species groups (*Eurotium herbariorum*, *Mucor circinelloides*, *Mucor hiemalis*, *Mucor plumbeus*, *Paecilomyces varioti*) were removed due a low number of spectra in each group. The removal of species groups with low number of spectra was necessary since in some situations a certain QT discarded almost all spectra for these species. In this situation it was hard to compare different QT approaches against each other. Thus, the reduced set used for the classification contained 14 species, corresponding to 913 spectra. This set was split: four runs were used as a calibration set, the other two runs were used as an independent validation set. The sets contained 573 and 340 spectra, respectively, before any QT was applied.

Spectra were pre-processed by (1) taking the first derivative by the Savitzky-Golay [20] algorithm with third order polynomial and a window size 9; (2) selection of the ranges [3050; 2800] & [1800; 900] cm^{-1} as informative for the analysis; (3) extended multiplicative signal correction (EMSC) [19] with linear and quadratic terms. The pre-processing was done separately for the calibration and the validation sets. The EMSC model established for the calibration set was applied to correct the spectra of the validation set – as it is supposed to be done in real case scenario.

The quality of spectra in the calibration set was evaluated using standard OPUS test, optimized versions of OPUS QT, and proposed EMSC QT. Spectra of the calibration set that did not pass the respective QT were discarded from the analysis. The standard OPUS QT was applied to the validation set in order to be able to compare different QT approaches. Thus, the validation was done on exactly the same data set. This allowed us testing the hypothesis that an optimized QT for the calibration data can improve the classification model.

2.1.2. FTIR spectra of yeasts

A set of 91 yeast strains from 13 different genera obtained from international yeasts collections were measured and analyzed by FTIR spectroscopy in a study performed by Shapaval et al. [3] The yeasts data set was organized in a taxonomic tree with 4 levels of phylogenetic hierarchy (phylum, class, family, and genus). The details about the taxonomic tree are provided in Supplementary materials (**Figure S2**). The classification problem for the yeasts data set was to identify genera.

The experimental design used for preparing yeasts spectra was the following: all strains were grown on five different growth media in six independent cultivation runs in a Bioscreen C cultivation system (Oy Growth Curves AB, Helsinki, Finland). In each run two micro-cultivations were performed in the Bioscreen C system. From each micro-cultivation, two technical replicates were obtained by FTIR spectroscopy. For the purposes of this study, the data corresponding to the growth medium Sabouraud broth (SAB) were selected. Sabouraud broth (SAB) is a standard growth medium used for detection, enumeration and identification of yeasts.

The whole data set (1943 spectra) was used to evaluate the quality of the data, to obtain OPUS parameter ranges, and to find optimal parameter thresholds. The data set was split into a calibration and a validation set: four runs were used for calibration, the other two runs were used for validation. The calibration and validation sets contained 1305 and 638 spectra, respectively, before any spectral QT was applied. The same spectral pre-processing was done as for the filamentous fungi spectra and the same procedure was used to compare QT results as described above.

2.1.3. FTIR spectra of bacteria

A set of 45 strains of Antarctic bacteria from 9 genera and 19 species was measured by FTIR spectroscopy in a study by M. Smirnova et al.[18] Bacteria data were organized in a taxonomic tree with 5 levels of phylogenetic hierarchy (phylum, class, family, genus, and species). The details about the taxonomic tree are provided in Supplementary materials (**Figure S3**). The classification problem for this data set was to identify species.

The experimental design, used for preparing bacteria spectra, was the following: all bacteria strains were cultivated on brain heart infusion (BHI) agar medium in three biological replicates – each performed on a separate day.

The whole data set (398 spectra) was used to evaluate the quality of data, to obtain OPUS parameter ranges, and to find optimal parameter thresholds. For the classification analysis, six species had to be removed (*Arthrobacter oryzae*, *Leifsonia kafniensis*, *Polaromonas glacialis*, *Pseudomonas extremaustralis*, *Pseudomonas weihenstephanensis*, *Psychrobacter glaciei*) due to the insufficient number of spectra to perform the analysis. Thus, the remaining data set contained 344 spectra of 8 genera and 13 species. The data set was split: two runs were used as a calibration set, one run was used as a validation set. The sets contained 228 and 116 spectra, respectively, before any QT was applied. The pre-processing was applied in the same manner with only one difference: the second derivative was calculated using Savitzky-Golay with second order polynomial and a window size 5. It is known that different parameters of Savitzky-Golay algorithm may be required to pre-process spectra. [46] These selected parameters of Savitzky-Golay algorithm have shown to be optimal for the classification results among other combinations tested (results not shown). The same procedure is used to compare QT results as suggested for other data sets.

2.1.4. FTIR images of filamentous fungal hyphae

7 pairs of microscopy images and FTIR images of filamentous fungal hyphae were analyzed in this study. FTIR images of hyphae were obtained from oleaginous filamentous fungus *Mucor circinelloides*, grown under lipid accumulation conditions – access of glucose and nitrogen limitation. Data matrix (128x128x765) contained 16384 spectra in total.

2.2. Classification methods

Multivariate classification analysis was used to assess the QTs performance. In order to compare the different QT approaches, we applied two methods that are frequently used to establish classifiers based on infrared spectra of biological materials, namely partial least squares regression (PLSR) [36] and Random

Forest (RF). [37] PLSR is a chemometric method based on latent variables [47], while RF is a tree-based method, both are widely used for classification of infrared spectra.

To establish a PLSR classifier, we apply a taxonomic PLSR where a PLSR classifier is established at each node of the taxonomic tree using only samples of the calibration set that are relevant for a given node. [12, 13] A PLSR classifier is established by regressing the matrix of indicator variables **Y** onto the matrix of FTIR spectra **X**, the method is also known as Partial Least Squares Discriminant Analysis (PLSDA). [48] Similar to Principal Component Analysis (PCA), the PLS algorithm finds new components represented by so-called PLS components for **X** and **Y**. For each component, the PLS model maximizes the residual co-variance matrix of **X** and **Y**. [49] An important parameter to be optimized in PLS modelling is the number of PLS components in **X** and **Y**. This is done by leave-one-run-out cross-validation. The optimal number of PLS components is determined as the smallest number which does not yield a significantly higher misclassification rate (MCR) than the number of PLS components corresponding to the minimum MCR, while the statistical significance of this difference was evaluated.

To predict classes, every sample follows the tree from top to bottom (see Supplementary materials for trees specifications Figure S1, S2 and S3) being classified by the corresponding classifiers. For more details on the method's set-up see Tafintseva et al. [12]

RF is a versatile tool used widely in classification analysis. It works well on a big number of classes without any need for hierarchical separation of classes. The method is very robust due to its nature: it builds-up an ensemble of trees with low correlations to each other and thus avoiding over-fitting of the model. Each tree is built up by using a random selection of samples from the original data set using bootstrapping (random sampling with replacement) where about two-third of all samples, the so-called "bootstrap" set, are selected for training and the remaining one-third, the so-called "out-of-bag" (OOB) set, is used for testing. Each node in a tree is optimized using a random subset of variables. A Gini impurity is used to define a variable used for splitting in each node, which is defined as follows: $i(\tau) = 1 - \sum_{k=1}^{K} p_k(\tau)^2$, where $p_k(\tau)$ is the fraction of samples belonging to the k^{th} group out of the total number of samples at node τ and K is the total number of groups considered. [50] To avoid overfitting the models, pruning of the trees or merging of leaves are common measures. To validate the model's performance, samples of the OOB test set are run through every tree and the majority voting defines the class' belongingness.

The number of trees in RF was optimized for each set between 100 and 500. The optimal number of trees for the filamentous fungi, yeasts and bacteria data was found to be 300, 300, and 250 trees, respectively. In total, 35, 35 and 25 variables (the square root of the total number of variables: $\sqrt{1195}$, $\sqrt{1195}$, and $\sqrt{597}$) were randomly selected for the optimization of nodes in each tree of the RF for filamentous fungi, yeasts and bacteria, respectively. Since pruning of the trees and leaves merging is not recommended for bagged trees, neither of the two was performed.

Data analyses were performed by standard algorithms, algorithms developed in-house, and open-source algorithms in Matlab, R2018a (The Mathworks Inc., Natick, USA).

2.3. Quality tests

2.3.1. Standard OPUS QT

The OPUS QT is a frequently used QT for infrared spectra. [11, 15, 26, 35, 51, 52] It assesses the quality of infrared spectra with regard to absorbance values, signal-to-noise ratio and intensity of the water vapor lines. Spectra that do not pass the test are considered to be of poor quality and usually discarded from analysis. [32, 34]

The parameters of standard OPUS QT are described in Table 1. Parameters of the QT are written in italic throughout the paper.

Table 1. OPUS quality test parameters as provided by Bruker.

Quality test parameters	min	max
Abs (X-range 1: 2100 – 1600 cm^{-1})	0.345000	1.245000
Noise (X-range 4: 2100 – 2000 cm^{-1})	0.000000	0.000150
S_1/N (X-range 2: 1700 – 1600 cm^{-1})	200.000000	0.000000
S_2/N (X-range 3: 1200 – 960 cm^{-1})	40.000000	0.000000
Water vapor (X-range 5: 1847 – 1837 cm^{-1})	0.000000	0.000300
S_1/W (X-range 2: 1700 – 1600 cm^{-1})	100.000000	0.000000
S_2/W (X-range 3: 1200 – 960 cm^{-1})	20.000000	0.000000

X-Range 1: 2100 – 1600 cm^{-1}. This range determines *Abs* parameter by calculating the difference between the maximum and minimum absorbance values of the original spectrum. For the spectra of satisfactory quality, the *Abs* parameter has to be higher than the min and lower than the max entry field corresponding to 0.345 to 1.245, respectively.

X-Range 2: 1700 – 1600 cm^{-1}. The range represents one of the characteristic spectral regions, namely an amide I, with the peak at 1650 cm^{-1}. The maximum and minimum values of the first derivative are calculated and the difference between these two values results in S_1.

The signal-to-noise ratio, S_1/N, is calculated dividing S_1 by the noise determined as explained in the X-Range 4. The minimum S_1/N value is equal to 200.0.

S_1 is also divided by the water vapor signal determined in the X-Range 5. The minimum S_1/W equals to 100.0.

X-Range 3: 1200– 960cm^{-1}. The range represents another characteristic spectral region related to polysaccharide ring vibrations. The maximum and minimum values of the first derivative is calculated and the difference between these two values results in S_2.

The signal-to-noise ratio, S_2/N, is calculated dividing S_2 by the noise determined as explained in the X-Range 4. The minimum S_2/N value is equal to 40.0.

S_2 is also divided by the water vapor signal determined in the X-Range 5. The minimum S_2/W equals to 20.0.

X-Range 4: 2100 – 2000 cm^{-1}. The range shows no absorbance bands and can therefore be used to estimate *Noise* parameter. The difference between the maximum and minimum values of the first derivative is calculated and is limited to *Noise* equal to 1.5 x 10^{-4}.

X-Range 5: 1847 – 1837 cm^{-1}. The range indicates strong water vapor absorbance but no sample absorbance and is therefore used to calculate water vapor. The difference between the maximum and minimum values of the first derivative is calculated and is limited by max water equal to 3 x 10^{-4}.

Spectral quality of each data set was assessed using quality analysis based on the standard OPUS QT. To do so, each full data set was quality analyzed by the OPUS QT. Relevant parameters and their ranges were obtained by calculating OPUS QT parameters for each spectrum of the considered data set. By screening through the parameter values, parameter ranges were identified and used for optimization of the thresholds of OPUS test parameters for each data set separately.

2.3.2. Alternative OPUS QT

We optimized thresholds of the OPUS QT for each microorganism individually. First, relevant QT parameters for each data set were selected. Second, optimal thresholds were obtained for each parameter. To find an optimal threshold for an OPUS test parameter, a simple grid search through wider parameter's ranges was performed according to the following procedure: 1) update the threshold of an OPUS QT parameter; 2) apply OPUS QT with the updated parameter to the calibration set; 2) establish a new calibration model using the spectra that passed the updated QT; 3) use the model to obtain predictions for spectra of the validation set that were quality tested according to the standard OPUS QT; 4) compare success rates for the validation (SR$_{val}$) with SR$_{val}$ of the previous model. Each time a parameter was optimized, other OPUS QT parameters were kept equal to their thresholds of the original standard OPUS QT. Since for each data set there were only few parameters which had to be optimized, there was no need for cross-optimization of the parameters.

2.3.3. EMSC QT

Let us briefly remind the idea underlying EMSC method. An EMSC model is described by

$$Z_{app}(\tilde{\nu}) = c + bZ_{ref}(\tilde{\nu}) + d\tilde{\nu} + e\tilde{\nu}^2 + \varepsilon$$

where Z_{app} is a measured spectrum, Z_{ref} is a reference spectrum, b is a multiplicative parameter, c, d, e are constant, linear and quadratic parameters, respectively, ε is a residual term, $\tilde{\nu}$ are spectral wavenumbers. The reference spectrum in the EMSC model is usually calculated as the average spectrum of the calibration set.

In order to set up an EMSC QT, we established an EMSC model for a data set and used the multiplicative parameter b of the model. The parameter b of the EMSC model correlates with the effective optical path length and is scaled by the total absorbance of the reference spectrum in the EMSC model. In order to make the multiplicative parameter *b* independent of the scaling of the reference spectrum and other effects such as the baseline variations in the reference spectrum, we performed baseline correction of the original spectra and normalized the reference spectrum. The normalization was done by the highest absorption peak in the data set at hand since different types of microorganisms exhibit different maximum absorption bands (**Figure 5**). We used the 1036 cm^{-1} band for filamentous fungi, and the 1650 cm^{-1} band for yeasts and bacteria. The corrected reference spectrum was then used to establish an EMSC model and spectra that had lower or higher *b* than the corresponding thresholds, *minb* and *maxb*, were removed as poor quality spectra.

2.4. EMSC binary segmentation (masks)

To obtain optimal EMSC binary segmentation or a mask we had to compare it to a binary segmentation of microscopy image. Since the microscopy image and the FTIR images are acquired simultaneously and therefore aligned, there was no need for any image registration. To obtain a binary segmentation of the microscopy image, manual annotation was done. To construct binary segmentation for FTIR images the following steps were done: 1) important spectral regions were selected [3900; 2600] & [2000; 900] cm^{-1}, 2) an EMSC model with linear and quadratic terms was established using spectra of all pixels with the selected spectral regions, 3) the *b* parameter of the EMSC model was used to establish masks: spectra with *b* values below a certain threshold were assigned to background pixels of the image. A threshold for *b* in a mask was defined in optimization, maximizing the Dice-Sørensen similarity coefficient [53, 54] known also as F_1 score between the EMSC mask and the annotated microscopy image mask. The coefficient allows estimating similarity between segmented objects in binary images. A similarity coefficient equal to 1 means a perfect match between two images.

3. Results

3.1. FTIR spectra of filamentous fungi

The standard OPUS QT was applied to the whole data set (1029 spectra). In total 463 spectra (45%) were identified as poor quality: 260 spectra (25.3%) had low *Abs*, 200 spectra (19.4%) had low S_1/N, 107 spectra (10.4%) had high *Noise*, 29 spectra (2.8%) had high *Abs*, 25 spectra (2.4%) had low S_1/W, and 1 (0.1%) spectrum had high *Water*. No spectra were identified as of low S_2/N. Among these poor quality spectra we could see that only 56 were both low *Abs* (out of 260) and low S_1/N (out of 200), and only 61 spectra were both low in S_1/N (out of 200) and high in *Noise* (out of 107). Most of the spectra (21 out of 29) with high *Abs* were also high in *Noise*. Further, 13 out of 25 spectra with low S_1/W were also low in *Abs*. Based on these observations, the following parameters were selected as the most relevant for OPUS QT optimization: *minAbs*, S_1/N, and *Noise*. The results of OPUS QT applied to the entire data set are summarized in Table 2 and spectra are presented in **Figure S4** in Supplementary materials.

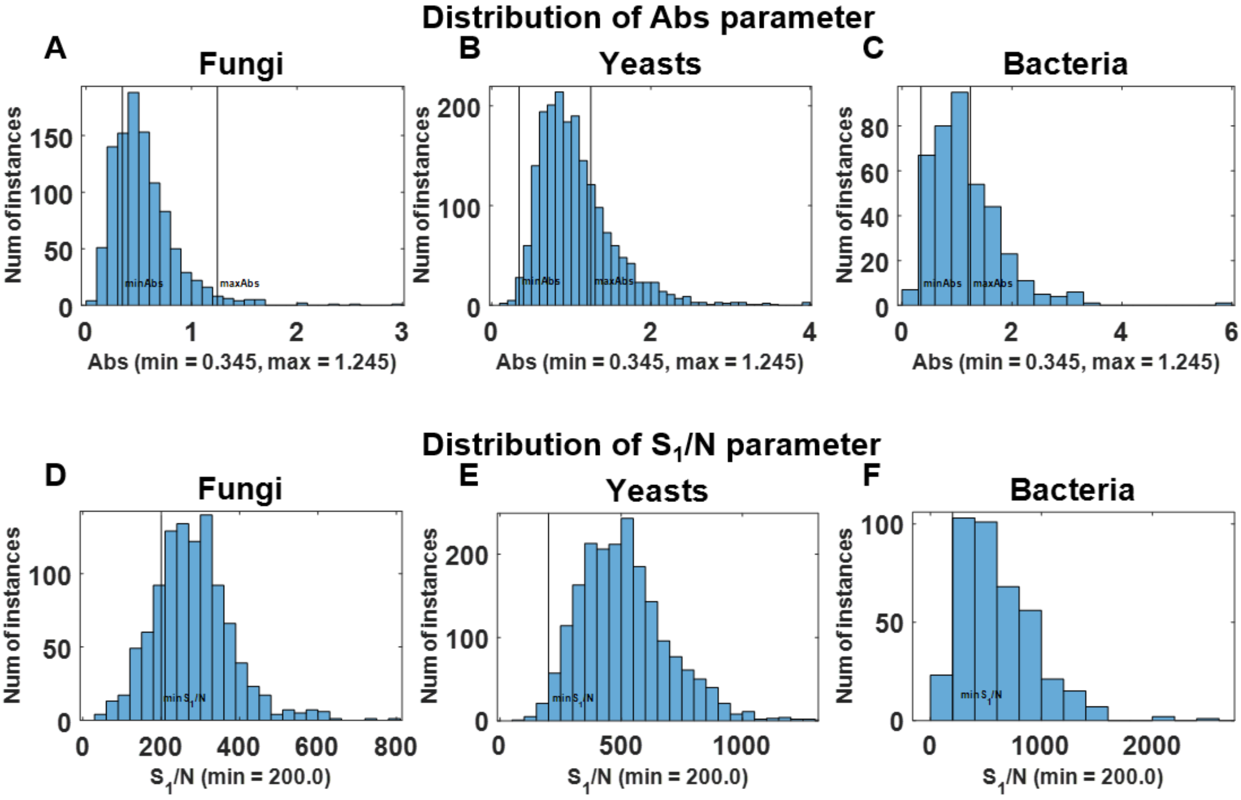

Figure 1. Distribution of *Absorbance (Abs)* parameter (upper panel) and S_1/N (lower panel) from OPUS quality test (QT) applied to the filamentous fungi (first column, A, D), yeasts (second column, B, E) and bacteria (third column, C, F) data sets. Red lines correspond to standard OPUS QT parameter thresholds for *Abs*: *minAbs* = 0.345, *maxAbs* = 1.245 and for S_1/N: S_1/N = 200.0. Spectra that are to the left of the *minAbs* and to the right of *maxAbs* in the upper panel and to the left of the $minS_1/N$ in the lower panel are to be discarded by OPUS QT.

Table 2. Selected as relevant OPUS quality test (QT) parameters of fungi, yeasts and bacteria data sets. Each data set is OPUS quality tested and the number of spectra that are identified as poor is provided together with the percent of total. Each data set size is provided in the first column.

Type (Total spectra)	Low *Abs*, <0.345	High *Abs*, >1.245	High *Noise*, >$1.5x10^{-4}$	Low S_1/N, <50	Low S_1/W, <100	High *Water*, > $3x10^{-4}$	Low S_2/N, <40
Fungi (1029)	260 25.3%	29 2.8%	107 10.4%	200 19.4%	25 2.4%	1 0.1%	0
Yeasts (1943)	18 0.9%	518 26.7%	127 6.5%	27 1.4%	0	0	0
Bacteria (398)	10 2.5%	138 34.7%	193 48.5%	23 5.8%	0	1 0.3%	5 1.3%

In order to compare classification results after different alternatives of OPUS QT were applied to the calibration data, the same standard OPUS quality test was applied to the validation set. For the filamentous fungi, 230 spectra out of 340 of the validation set passed the OPUS QT and were used for the validation. The classification results are summarized in Table 3 in the form of SR and presented graphically in **Figure 2** as MCR (MCR=100-SR, %). When applying the standard OPUS QT to the calibration set (573 spectra), only 304 spectra passed the test, i.e. almost half of the spectra were discarded because of poor quality. The corresponding PLSR and RF models yielded success rates of validation SR_{val}=67.0% and 79.6%, respectively. For the same data set without any QT we established and obtained PLSR and RF models with success rates of validation SR_{val}=75.2% and SR_{val}=83.5%, respectively. These results indicate that the OPUS QT removes a lot of spectra that could potentially improve the predictive performance of the classification models. In order to investigate this further, we systematically changed the thresholds of selected parameters of the OPUS QT in order to optimize them. We searched for optimal thresholds of the three parameters, *minAbs*, S_1/N, and *Noise*, that had shown to be responsible for discarding the lion's share of the spectra (see **Figure 1A, D**, and **Figure S10, S11** in Supplementary materials). The *minAbs* parameter was optimized in the range [0.1; 0.345] (see Figure 1A), while *maxAbs* was kept equal to the standard OPUS QT value *maxAbs*=1.245. The optimal values identified for PLSR and RF were $minAbs_{opt}$=0.2 with SR_{val}=81.3% and $minAbs_{opt}$=0.3 with SR_{val}=84.8%, respectively. When tuning the *Noise* parameter in the range [$1.5x10^{-4}$; $2.2x10^{-4}$] (see Figure S10) we observed that optimal threshold values $Noise_{opt}$=1.8x10^{-4} with a success rate of SR_{val}=73.9% and $Noise_{opt}$=1.7x10^{-4} with SR_{val}=80.4% for PLSR and RF, respectively. The search for optimal S_1/N in the range [50; 200] (see Figure 1D) showed different results for PLSR and RF models: S_1/N_{opt}=150 with SR_{val}=71.3% for PLSR and S_1/N_{opt}=50 with SR_{val}=80.9% for RF. Thus, we observe that tuning OPUS QT parameters can improve classification models considerably.

In the following we present the results of an alternative approach for spectral QT based on EMSC using the multiplicative parameter *b* as a QT parameter. The thresholds corresponding to the OPUS standard *Abs* parameter could also be considered as standard for parameter *b: minb*=0.345, *maxb*=1.245. These thresholds in EMSC QT applied to the whole data set resulted in 121 spectra (11.8%) with low *b* and 100 spectra (9.7%) with high *b* to be discarded. The results of EMSC QT are presented in **Figure S5** in Supplementary materials. When applying this QT to the calibration data we obtained models with SR_{val}=71.7% for PLSR and SR_{val}=80.9% for RF. Further, the threshold of *b* parameter was optimized in a range [0.1; 2] since the spectra at hand had *b* parameter in the range (see **Figure S12**). The optimal results were obtained for $minb_{opt}$=0.2, $maxb_{opt}$=1.5 with SR=82.6% for PLSR and $minb_{opt}$=0.1, $maxb_{opt}$=1.7 with SR_{val}=84.3% for RF. We see that optimizing the thresholds for the parameter *b* resulted clearly in the best classification result for both methods. However, we can see that the optimal parameter values vary a bit in both cases. The misclassification rate of validation MCR_{val} for PLSR and RF applying all considered quality tests are presented together in **Figure 2A**.

It is obvious that we can expect correlations between the quality parameters *Abs*, S_1/N, *Noise* and the EMSC parameter *b*. The results of cross-correlation analysis are presented in Table S1 in Supplementary materials. We can see that there is a correlation between *Abs* and *b* parameters R^2(*Abs*, *b*) = 0.81 and a slight correlation between *Abs* and *Noise* parameters R^2(*Abs*, *Noise*) = 0.50, and otherwise there is no correlation between other parameters. Comparing the spectra removed by QTs (see Figure S4, S5), we can see that in total 126 spectra were identified as poor quality by EMSC QT with standard thresholds for *b* parameter and 269 spectra by OPUS QT. Almost all spectra (114 out of 126) removed by EMSC QT were also identified by standard OPUS QT as poor quality spectra. A lot more spectra (155 spectra) were identified as poor quality by OPUS and not by EMSC QT. Comparing classification results using the quality tested data we can conclude that EMSC QT performs better quality control of the data. EMSC QT discards all poor quality spectra which hinder establishing good prediction models, whereas OPUS QT tends to discard many more spectra, part of which contribute to establishing good models.

Table 3. Validation results of different QT approaches applied to filamentous fungi data set when using PLSR and RF classification methods.

Spectral quality test	PLSR			RF		
	Parameters	Spectra[1]	SR_{val}[2], %	Parameters	Spectra	SR_{val}, %
No quality test	-	573	75.2	-	573	83.5
Standard OPUS QT	See Table 1	304	67.0	See Table 1	304	79.6
Alternative OPUS QT Opt *minAbs*	*minAbs*=0.2 *maxAbs*=1.245	420	81.3	*minAbs*=0.3 *maxAbs*=1.245	335	84.8
Alternative OPUS QT Opt *Noise*	*Noise*=$1.8x10^{-4}$	313	73.9	*Noise*=$1.7x10^{-4}$	310	80.4
Alternative OPUS QT Opt S_1/N	S_1/N=150	337	71.3	S_1/N=50	341	80.9
Standard EMSC QT	*minb*=0.345 *maxb*=1.245	447	71.7	*minb*=0.345 *maxb*=1.245	447	80.9
Opt EMSC QT	*minb*=0.2 *maxb*=1.5	541	82.6	*minb*=0.1 *maxb*=1.7	556	84.3

[1] Number of spectra left after QT applied to a calibration set. The total number of spectra in the original calibration set was 573.

[2] Success rate of classification on the external validation set. The same QT was applied to the validation set for easier comparison of the results.

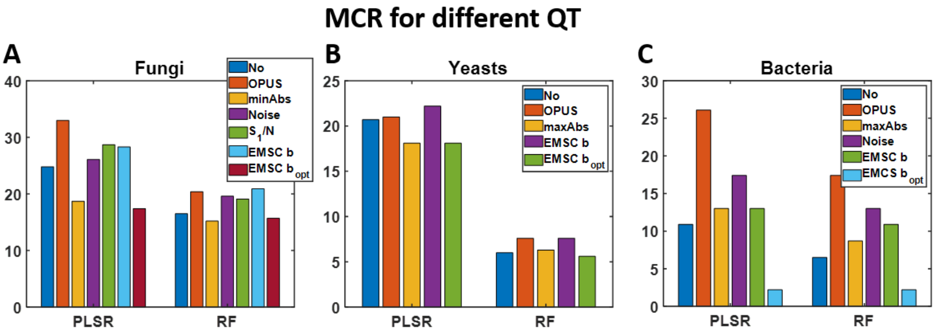

Figure 2. Misclassification (error) rate (MCR, %) of validation on an independent test set for filamentous fungi (A), yeasts (B) and bacteria (C). Standard OPUS QT was applied to the validation set while different QTs were applied to the calibration set: standard OPUS QT (in red) and different alternatives of OPUS QT, optimizing *minAbs*, *maxAbs*, *Noise*, and *S_1/N*, EMSC QT was applied with standard *b* parameter range [0.345; 1.245] and optimized *b* corresponding to EMSC b_{opt}. The result of classification when no quality test is applied is presented in blue.

3.2. FTIR spectra of yeasts

The whole data set (1943 spectra) was quality tested by the standard OPUS QT. In total 571 spectra (29.4%) were identified as poor quality spectra. The majority of them, 518 spectra (26.7%) had high *Abs*, 127 spectra (6.5%) had high *Noise*, 27 spectra (1.4%) had low *S_1/N*, and only 18 spectra (0.9%) had low *Abs*. No spectra were identified as of low *S_1/W*, low *S_2/N* or high *Water*. Among these poor quality spectra we could see that almost all high *Noise* spectra (111 out of 127) were also high *Abs*. This allowed us to conclude that yeasts spectra were generally high in *Abs* and therefore *maxAbs* parameter was selected as the only relevant for OPUS QT optimization. The results of the OPUS QT are summarized in Table 2 and spectra are presented in **Figure S6** in Supplementary materials.

Applying the EMSC QT to the whole data set with the thresholds for *b: minb*=0.345 and *maxb*=1.245, resulted in 14 (0.7%) spectra with low *b* parameter and 447 (23%) spectra with high *b* parameter to be discarded (see **Figure S7** in Supplementary materials).

For classification analysis, 463 spectra out of 638 of the validation set that passed the OPUS QT were used for the validation. The results of the validation when different spectral QTs were applied to the calibration set are summarized in Table 4. Applying standard OPUS QT to the calibration set left us with only 909 spectra out of the 1305 from the original set. After applying the OPUS QT, PLSR and RF models were established with SR_{val}=79.0% and 92.4% for the validation, respectively. In the case when not applying any QT, slightly better models were obtained with both methods: SR_{val}=79.3% for PLSR and SR_{val}=94.0% for RF.

Since *maxAbs* was almost the only major filter in the OPUS QT for the yeasts data set (see **Figure 1B, E**, and **Figure S13, S14** in Supplementary materials), we optimized it in a range [1.3; 2.5] (see Figure 1B), while *minAbs* was kept equal to the standard OPUS value *minAbs*=0.345. The best models were obtained with the same parameter value $maxAbs_{opt}$=2.2 and SR_{val}=81.9% for PLSR and SR_{val}=93.7% for RF.

An EMSC QT with thresholds *minb*=0.345 and *maxb*=1.245 for *b* parameter was applied to the calibration data set and the following results were obtained: SR_{val}=77.8% for PLSR and SR_{val}=92.4% for RF. This rather weak classification results suggests that the selected parameter range is not optimal for the given yeasts data set. Further the parameter of EMSC QT were optimized in a range [0.3; 2.2] (see **Figure S15**) and the best models were obtained for PLSR with $minb_{opt}$=0.4, $maxb_{opt}$=1.8 and SR_{val}=81.9% and for RF with $minb_{opt}$=0.4 and $maxb_{opt}$=2.2 and SR_{val}=94.4%. The summary of the results but in a form of error MCR_{val} (%) are presented in **Figure 2B**.

The results of cross-correlation showed that *Abs* and *b* parameters are correlated, R^2(*Abs*, *b*) = 0.88 (see Table S2 in Supplementary materials). Closer look at the spectra (see Figure S6, S7) shows that in total 326 spectra were identified as poor quality by EMSC QT with standard threshold values for *b* parameter against 396 identified by OPUS QT. Of these 298 spectra were identified as poor quality spectra by both QTs. Classification results suggest that the performance of the two QTs are similar, however OPUS QT discarded more spectra.

Table 4. Validation results of different QT approaches applied to yeasts data set when using PLSR and RF classification methods.

Quality test	PLSR			RF		
	Parameters	Spectra[1]	SR_{val}[2], %	Parameters	Spectra	SR_{val}, %
No quality test	-	1305	79.3	-	1305	94.0
Standard OPUS QT	See Table 1	909	79.0	See Table 1	909	92.4
Alternative OPUS QT Opt *maxAbs*	*minAbs*=0.345 *maxAbs*=2.2	1188	81.9	*minAbs*=0.345 *maxAbs*=2.2	1188	93.7
Standard EMSC QT	*minb*=0.345 *maxb*=1.245	979	77.8	*minb*=0.345 *maxb*=1.245	979	92.4
Opt EMSC QT	*minb*=0.4	1196	81.9	*minb*=0.4	1251	94.4

	maxb=1.8			*maxb*=2.2		

[1] Number of spectra left after QT applied to a calibration set. The total number of spectra in the original calibration set is 1305.

[2] Success rate of classification on the external validation set. The same QT was applied to the validation set for easier comparison of the results.

3.3. FTIR spectra of bacteria

The whole data set (398 spectra) was quality tested by standard OPUS QT. More than half of the spectra, in total 227 spectra (57%), were identified as poor quality spectra. The majority of them, 193 spectra (48.5%), had high *Noise*, 138 spectra (34.7%) had high *Abs*, 23 spectra (5.8%) had low S_1/N, only 10 spectra (2.5%) had low *Abs*, 5 (1.3%) had low S_2/N, and 1 (0.3%) had high *Water* value. Among these poor quality spectra almost all high *Abs* spectra (115 out of 138) were also high in *Noise*. Almost all low S_1/N spectra (22 out of 23) were high in *Noise*. These observations allowed us to conclude that the spectra of bacteria were high in *Abs* and *Noise*. Therefore *maxAbs* and *Noise* parameter were selected as the most relevant for OPUS QT optimization (see **Figure 1C, F**, and **Figure S16, S17** in Supplementary materials). A summary of OPUS QT results is provided in Table 2 while spectra are presented in **Figure S8** in Supplementary materials.

Applying EMSC QT to the entire data set with the standard parameter range *b* in [0.345; 1.245] had low *b* for 14 (3.5%) and high *b* parameter for 144 (36.2%) spectra (see **Figure S9**).

For classification analysis, the validation set was OPUS quality tested and 46 spectra out of 116 that passed the quality test were used further in validation. The results of the validation when the different quality tests were applied to the calibration set are summarized in Table 5. Applying the standard OPUS QT to the calibration set left us with only 110 spectra out of the 228 from the original set. Using this quality tested calibration set, PLSR and RF models were established with SR_{val}=73.9% and 82.6% for the validation, respectively. In the case of not applying any QT, much better models were obtained with both classification methods: SR_{val}=89.1% for PLSR and SR_{val}=93.5% for RF. This is most likely due to the amount of the spectra used to establish calibration models: more than half of the spectra of the calibration set are discarded by the OPUS QT.

To improve OPUS QT performance, parameter *maxAbs* was optimized in a range [1.2; 2.4] (see Figure 1C), while *minAbs* was kept equal to the standard OPUS value *minAbs*=0.345. The best models were obtained with $maxAbs_{opt}$=1.8 and SR_{val}=87.0% for PLSR and $maxAbs_{opt}$=2.1 and SR_{val}=91.3% for RF. *Noise* was optimized in a range [1.5; 4.5] x10^{-4} (see Figure S16). The best models were obtained with $Noise_{opt}$=2 x10^{-4} and SR_{val}=82.6% for PLSR and $Noise_{opt}$=3.5 x10^{-4} SR_{val}=87.0% for RF.

The EMSC QT with the threshold *minb*=0.345 and *maxb*=1.245 for the *b* parameter was applied and the following results were obtained: SR_{val}=87% for PLSR and SR_{val}=89.1% for RF. Furthermore, the parameter of the EMSC QT was optimized in the range [0.3; 3] (see **Figure S18**) and the best models were obtained with the same success rate SR_{val}=97.8% with $minb_{opt}$=0.4, $maxb_{opt}$=2.1 for PLSR and $minb_{opt}$=0.5, $maxb_{opt}$=3 for RF. The summary of the results in a form of MCR_{val} (%) are presented together in **Figure 2C**.

The results of cross-correlation between *Abs*, *Noise* and *b* parameters showed that there is a correlation between *Abs* and *b* parameters, R^2(*Abs*, *b*) = 0.90, but otherwise other parameters are not correlated (see Table S3 in Supplementary materials). In total 80 spectra were discarded by EMSC QT with standard *b* parameter thresholds, whereas the standard OPUS QT discarded 118 spectra. All except 4 spectra identified by EMSC QT were identified by OPUS QT, 42 spectra were identified as poor quality by OPUS QT and not by EMSC. The classification results suggest that the performance of classification models is highly dependent on the results of quality analysis and the EMSC QT performs better quality control.

Table 5. Validation results of different QT approaches applied to bacteria data set when using PLSR and RF classification methods.

Quality test	PLSR			RF		
	Parameters	Spectra[1]	SR_{val}[2], %	Parameters	Spectra	SR_{val}, %
No quality test	-	228	89.1	-	228	93.5
Standard OPUS QT	See Table 1	110	73.9	See Table 1	110	82.6
Alternative OPUS QT Opt *maxAbs*	*minAbs*=0.345 *maxAbs*=1.8	117	87.0	*minAbs*=0.345	119	91.3

				maxAbs=2.1		
Alternative OPUS QT Opt *Noise*	*Noise*=2 x10^{-4}	127	82.6	*Noise*=3.5 x10^{-4}	142	87.0
Standard EMSC QT	*minb*=0.345 *maxb*=1.245	148	87.0	*minb*=0.345 *maxb*=1.245	148	89.1
Opt EMSC QT	*minb*=0.4 *maxb*=2.1	199	97.8	*minb*=0.5 *maxb*=3	189	97.8

[1] Number of spectra left after QT applied to a calibration set. The total number of spectra in the original calibration set is 228.

[2] Success rate of classification on the external validation set. The same QT was applied to the validation set for easier comparison of the results.

3.4. FTIR images of filamentous fungal hyphae

Finally, we investigated if EMSC parameters can also be used for quality control and binary segmentation or masking of FTIR images. In **Figure 4A** microscopy image and in **Figure 3A** sliced at 3010cm^{-1} FTIR image of fungal hyphae under the study are shown. To build a mask, an EMSC model was established using spectra of the whole FTIR image with the only spectral region of interest (see Materials and Methods). All parameters of the EMSC model are presented in **Figure 3B-3E**. As it could be seen from Figure 3D the multiplicative parameter *b* of the model is the one which strongly resembles the microscopy image. The residual after EMSC correction does not contain almost any information (see **Figure 3F**). Therefore, the *b* parameter is suggested for differentiating between background and foreground spectra. In this study parameter range for *b* in optimization was chosen to be [0.5; 2.5] based on the distribution of EMSC *b* parameter values of the entire image (see **Figure S19**). The manually annotated image of the microscopy image is presented in **Figure 4B** on top of the microscopy image. The Dice coefficient between two binary images was used to optimize *b* parameter for establishing the EMSC masks. The best result was obtained for b_{opt}=2.2 with the Dice coefficient equal to 0.89 and the mask is shown in **Figure 4D** on top of the microscopy image. Comparison of two masks is presented in **Figure 4C**. We can see that the mask obtained by EMSC segmentation is very similar to the manually annotated mask and cover well the sample from the image.

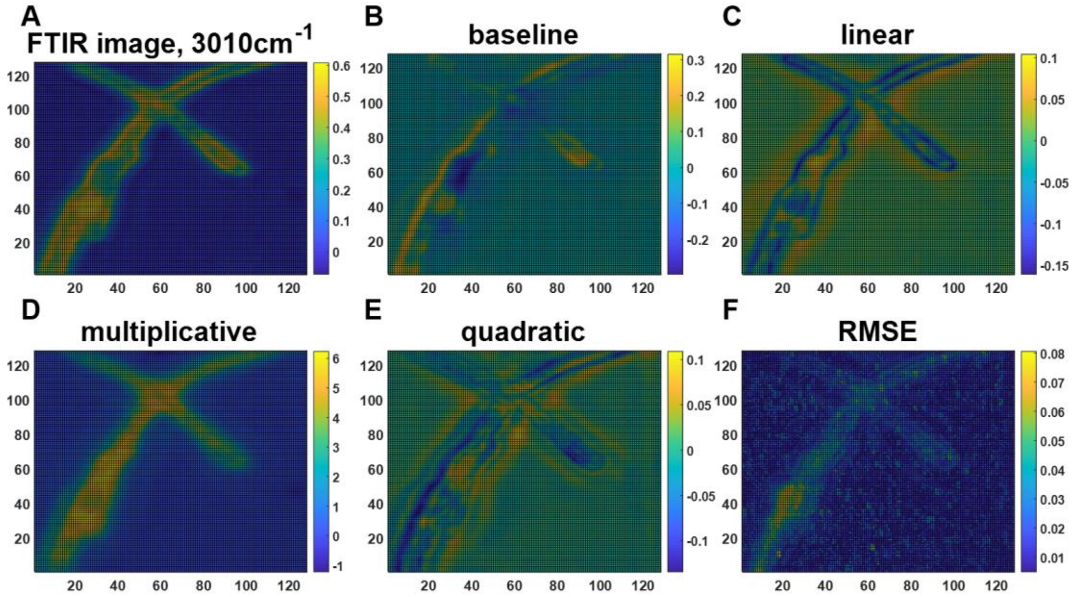

Figure 3. FTIR image slice at 3010cm^{-1} (A), parameters of EMSC model established on the whole fungal image (B-E), and root-mean-square-error after EMSC correction (F).

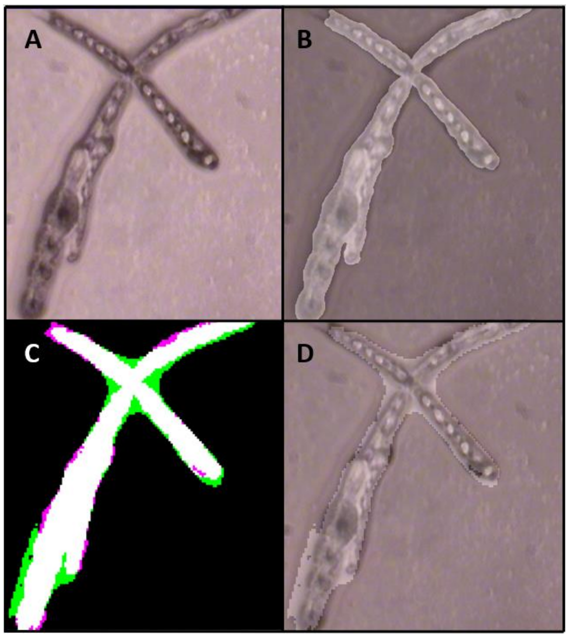

Figure 4. Microscopy image of fungal hyphae (A) and image segmentation results: by manually annotated mask (B) and EMSC mask (D) with *b*=2.2. Comparison of two masks are shown (C) where white is a common area identified as sample by both masks, green is sample area identified by EMSC mask but not annotated and pink is annotated as sample but identified as background by the EMSC mask. Dice coefficient is equal to 0.89.

4. Discussion

OPUS QT was originally developed for spectra of bacteria and was generally accepted and used for quality check of spectra of different origin, while our study shows that spectra of other microorganisms have different spectral quality properties than bacteria spectra. For example, most of poor quality spectra for filamentous fungi were identified due to low *Abs* and low S_1/N, many yeast spectra had bad quality due to only high *Abs* while the OPUS QT test discarded many bacteria spectra due to high *Abs* and *Noise*. Therefore, one of the drawbacks of the standard OPUS QT is that it does not take into account the fact that spectra of different microorganisms have different spectral qualities.

In order to explain the differences in the spectral quality properties of filamentous fungi, yeasts and bacteria according to the OPUS QT, spectra of these microorganisms in this study were averaged and plotted together (see Figure 5). The differences in spectra of these microorganisms are clearly seen in different spectral regions. The whole fingerprint region 1500-500 cm^{-1} is quite specific for each type of microorganism. The strongest peak for yeasts and bacteria corresponds to amide I peak at 1650 cm^{-1} followed by amide II peak at 1550 cm^{-1} for bacteria and polysaccharide peak at 1080 cm^{-1} for yeasts. The average spectrum of filamentous fungi is very different from the average spectra of bacteria and yeasts with the strongest peak in the polysaccharide region at 1036 cm^{-1} followed by amide I peak at 1650 cm^{-1}.

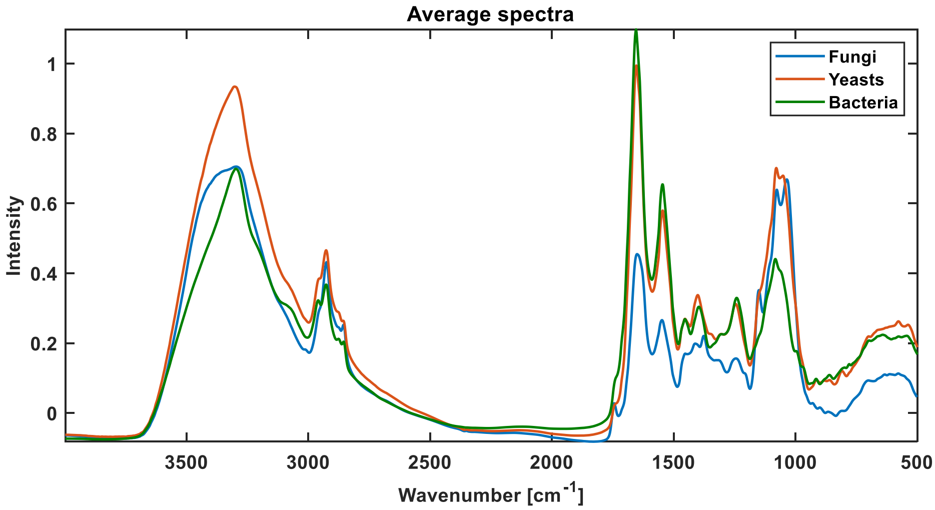

Figure 5. Averaged spectra of filamentous fungi, yeasts and bacteria of the corresponding data sets used in the study.

Thus, it becomes obvious why the majority of the poor quality spectra for filamentous fungi were identified based on low *Abs* values and low S_1/N (Table 2). Both of these parameters of the OPUS QT are calculated using the region of amide I peak which is not the strongest for the filamentous fungi spectra and therefore a large number of spectra was discarded. A suggestion for a QT of filamentous fungi spectra would be to lower down the threshold for the *minAbs* to 0.2. The results from the calibration modelling shows that such a change of the threshold in the QT allows increasing the amount of spectra that passes the test considerably, which had a positive effect on the established prediction models (see Table 3). Since the strongest peaks for filamentous fungi are polysaccharide peaks at 1036 cm^{-1} and 1080 cm^{-1} covered by S_2 of OPUS QT (S_2 is calculated on 1200 – 960 cm^{-1}), the parameter S_2/N is highly relevant for spectra of filamentous fungi. However, the threshold suggested in OPUS QT, S_2/N=40.0, is too low for this type of spectra. In a search for optimal S_2/N, the best results were obtained for S_2/N=220.0 with 278 spectra in the calibration set and SR_{val}=72.6% for the PLSR model, S_2/N=210.0 with 286 spectra in the calibration set and SR_{val}=81.3% for the RF model. Thus, it is suggested to increase the threshold to S_2/N=200.0 (see Figure S11 in Supplementary materials for the distribution of S_2/N parameter for fungi).

The majority of poor quality spectra for yeasts were selected based only on high *Abs*. The optimization of the OPUS QT showed that an increase of *maxAbs* to 2.2 yielded better models for PLSR and RF classifiers (see Table 4). From Figure 5 we can see that the amide I for yeasts is almost as high as for bacteria and polysaccharide peaks are as high as for filamentous fungi. Therefore, both S_1 and S_2 signals are important but need to be optimized. Optimizing S_1/N, the best results were obtained for S_1/N=220.0 with 901 spectra in the calibration set and SR_{val}=79.5% for the PLSR model and SR_{val}=93.1% for the RF model. Optimizing S_2/N, the best results were obtained for S_2/N=140.0 with 898 spectra in the calibration set and SR_{val}=80.3% for the PLSR model and S_2/N=130.0 with 906 spectra in the calibration set and SR_{val}=93.1% for the RF model. Therefore, we suggest increasing the thresholds of both parameter to S_1/N=220.0 and S_2/N=150.0 (see Figure 1E, Figure S14 for the distributions of the corresponding parameters).

Poor quality spectra for bacteria were determined by the OPUS QT due to high *Noise* and high *Abs*. The thresholds for these parameters are quite low for the type of spectra which explains why so many spectra are filtered out. An increase of the parameters to *maxAbs*=2 and *Noise*=4 x10^{-4} would reduce the amount of filtered out spectra and allow establishing better classification models (see Table 5 for the results, Figure 1C, and Figure S16 in Supplementary materials for the distribution of the parameters).

Another drawback of the OPUS QT is that it discards too many spectra. In this study 463 spectra (45% of the total number) were identified as of poor quality for filamentous fungi, 571 spectra (29.4%) for yeasts, and 227 (57%) for bacteria. These numbers are too high to be satisfied with the quality test. There are two reasons for that: 1) a lot of the discarded spectra are of good quality as can be seen from Figs.S4,S6,S8; 2) the results of classification analysis based on the OPUS quality tested data are a lot worse than those based on no quality tested data. At the same time when alternative optimized ranges of OPUS QT parameters were considered and thresholds were optimized, we could see that more spectra passed through the test and better classification models were obtained. However, best results of classification performance by both PLSR and RF methods were obtained on the data quality tested by EMSC QT, except for two cases: 1) two PLSR models with the same success rates SR=81.9% were established on the yeasts data that has passed an optimized OPUS QT and EMSC QT and 2) RF model was slightly better on the fungi data that has passed optimized OPUS QT with SR=84.8% against SR=84.3% for the model established on the data that has passed EMSC QT.

Thus, discarding spectra has to be always done with care since it is undesirable to remove any important variation in the data. This study suggests that EMSC QT is the best approach to quality test spectra.

Optimal b ranges to be used for the EMSC QT can be suggested based on results of classification analysis and distribution of b parameter for spectral data (Figure S12, S15, S18): b = [0.2, 1.5] for filamentous fungi, and b = [0.4, 2] for yeast and bacteria. Here we can remind again that the suggested thresholds are data set independent for each data set type (fungi, yeasts, bacteria) since normalization of the reference spectrum for each data set type is done by the highest peak of the spectra. This implies that in order to apply EMSC QT with the thresholds proposed in this study, one needs to normalize the reference spectrum of an established EMSC model in exactly same manner as it is done in this study (see Material and

Methods). Even though we could not obtain a very general QT approach for all types of spectral data in this study, instead we had to find *b* parameter thresholds for spectra of each type of microorganism separately, it is still much more general than the proposed approaches available in literature which are mostly based on signal-to-noise ratio in spectra, peak intensities and noise. [27-31]

Another interesting application of the multiplicative parameter *b* of the EMSC model is suggested in this study, namely binary segmentation or construction of masks for FTIR images. The idea follows naturally from the fact that *b* parameter representing the effective optical path length, correlates with the sample thickness. [42] This has been shown again in this study: the *b* parameter (Figure 3D) resembles well the morphology represented by the sample thickness in the FTIR image of filamentous fungi (Figure 4A) and thus can be used to find regions of interest in the image. To construct a mask, a threshold for *b* needs to be assigned such that spectra which are below the threshold are defined as a background and thus correspond to the mask. Successful results of the binary segmentation confirm the hypothesis that *b* parameter can be used for this purpose and masks that are close to manually annotated images can be obtained.

5. Conclusion

Many studies mention that infrared spectra are subjected to a quality test, however details about the quality tests are barely discussed in the scientific literature. This study presents a thorough evaluation of the commonly used OPUS QT applied to three infrared spectral datasets of filamentous fungi, yeasts, and bacteria. When examining the OPUS QT, we found that the suggested thresholds for the OPUS QT parameters are very strict. The OPUS QT in general removes too many spectra what results in poorer performances of subsequent classification analysis.

Thus, in order to achieve the best possible classification results for a given data set, we suggested to optimize the thresholds for the parameters *Abs*, *Noise*, S_1/N and S_2/N to make sure only spectra that weaken a classifier are removed. We show that optimizing ranges of OPUS QT can improve classification models results by PLSR and RF methods considerably. Further, our paper suggested a new EMSC QT to assess the quality of FTIR spectra, where the multiplicative EMSC *b* parameter was used as a QT parameter. We show that the EMSC QT results in the best classification models for all data sets investigated. While the potential of the EMSC QT was demonstrated with high-throughput infrared spectral data of filamentous fungi, yeasts and bacteria, it is generally applicable to any type of infrared or Raman spectral data. We show further that the *b* parameter of the EMSC model can be used for binary segmentation of FTIR images using images of filamentous fungi.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's website.

Matlab codes for the quality test and classification routines are available in https://gitlab.com/BioSpecNorway/codes-for-emsc-quality-test-paper-by-v.tafintseva.

References

1. Naumann, D., D. Helm, and H. Labischinski, *Microbiological characterizations by FT-IR spectroscopy.* Nature, 1991. **351**(6321): p. 81-82.
2. Amiel, C., et al., *FTIR spectroscopy and taxonomic purpose: Contribution to the classification of lactic acid bacteria.* Le Lait, 2001. **81**(1-2): p. 249-255.
3. Shapaval, V., et al., *FTIR spectroscopic characterization of differently cultivated food related yeasts.* Analyst, 2013. **138**(14): p. 4129-4138.
4. Wenning, M. and S. Scherer, *Identification of microorganisms by FTIR spectroscopy: perspectives and limitations of the method.* Applied microbiology and biotechnology, 2013. **97**(16): p. 7111-7120.
5. Shapaval, V., et al., *Fourier transform infrared spectroscopy for the prediction of fatty acid profiles in Mucor fungi grown in media with different carbon sources.* Microbial Cell Factories, 2014. **13**.
6. Kosa, G., et al., *Microtiter plate cultivation of oleaginous fungi and monitoring of lipogenesis by high-throughput FTIR spectroscopy.* Microbial cell factories, 2017. **16**(1): p. 101.
7. Forfang, K., et al., *FTIR spectroscopy for evaluation and monitoring of lipid extraction efficiency for oleaginous fungi.* PloS one, 2017. **12**(1): p. e0170611.
8. Kaminskyj, S., et al., *High spatial resolution analysis of fungal cell biochemistry–bridging the analytical gap using synchrotron FTIR spectromicroscopy.* FEMS microbiology letters, 2008. **284**(1): p. 1-8.

9. Szeghalmi, A., S. Kaminskyj, and K.M. Gough, *A synchrotron FTIR microspectroscopy investigation of fungal hyphae grown under optimal and stressed conditions.* Analytical bioanalytical chemistry, 2007. **387**(5): p. 1779-1789.
10. Shapaval, V., et al., *A high-throughput microcultivation protocol for FTIR spectroscopic characterization and identification of fungi.* Journal of biophotonics, 2010. **3**(8-9): p. 512-521.
11. Shapaval, V., et al., *Characterization of food spoilage fungi by FTIR spectroscopy.* Journal of Applied Microbiology, 2013. **114**(3): p. 788-796.
12. Tafintseva, V., et al., *Hierarchical classification of microorganisms based on high-dimensional phenotypic data.* J Biophotonics, 2018. **11**(3).
13. Liland, K.H., A. Kohler, and V. Shapaval, *Hot PLS-a framework for hierarchically ordered taxonomic classification by partial least squares.* Chemometrics and Intelligent Laboratory Systems, 2014. **138**: p. 41-47.
14. Kosa, G., et al., *High-throughput screening of Mucoromycota fungi for production of low-and high-value lipids.* Biotechnology for biofuels, 2018. **11**(1): p. 66.
15. Colabella, C., et al., *Merging FT-IR and NGS for simultaneous phenotypic and genotypic identification of pathogenic Candida species.* PloS one, 2017. **12**(12): p. e0188104.
16. Shapaval, V., et al., *Biochemical profiling, prediction of total lipid content and fatty acid profile in oleaginous yeasts by FTIR spectroscopy.* Biotechnology for biofuels, 2019. **12**(1): p. 140.
17. Kohler, A., et al., *High-throughput biochemical fingerprinting of Saccharomyces cerevisiae by Fourier transform infrared spectroscopy.* Plos One, 2015. **10**(2): p. e0118052.
18. Smirnova, M., et al., *Genotypic and phenotypic characteristics of bacteria isolated from the green snow of coastal area of Eastern part of* FEMS Microbiology Ecology 2019. **submitted**.
19. Martens, H. and E. Stark, *Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy.* J Pharm Biomed Anal, 1991. **9**(8): p. 625-35.
20. Savitzky, A. and M.J.E. Golay, *Smoothing and differentiation of data by simplified least squares procedures.* Anal. Chem., 1964. **36**: p. 1627-1639.
21. Barnes, R., M.S. Dhanoa, and S.J. Lister, *Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra.* Applied spectroscopy, 1989. **43**(5): p. 772-777.
22. Sun, J., *A correlation principal component regression analysis of NIR data.* Journal of Chemometrics, 1995. **9**(1): p. 21-29.
23. Kohler, A., et al., *Estimating and correcting Mie scattering in synchrotron-based microscopic Fourier transform infrared spectra by extended multiplicative signal correction.* Applied spectroscopy, 2008. **62**(3): p. 259-266.
24. Bassan, P., et al., *Resonant Mie scattering (RMieS) correction of infrared spectra from highly scattering biological samples.* Analyst, 2010. **135**(2): p. 268-277.
25. Konevskikh, T., et al., *Mie scatter corrections in single cell infrared microspectroscopy.* Faraday discussions, 2016. **187**: p. 235-257.
26. Oust, A., et al., *FT-IR spectroscopy for identification of closely related lactobacilli.* J Microbiol Methods, 2004. **59**(2): p. 149-62.
27. Lasch, P., et al., *Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis.* Biochimica et Biophysica Acta -Molecular Basis of Disease, 2004. **1688**(2): p. 176-186.

28. Kohler, A., et al., *Multivariate image analysis of a set of FTIR microspectroscopy images of aged bovine muscle tissue combining image and design information.* Analytical bioanalytical chemistry, 2007. **389**(4): p. 1143-1153.
29. Kneipp, J., et al., *Detection of pathological molecular alterations in scrapie-infected hamster brain by Fourier transform infrared (FT-IR) spectroscopy.* Biochimica et Biophysica Acta -Molecular Basis of Disease, 2000. **1501**(2-3): p. 189-199.
30. Kuepper, C., et al., *Label-free classification of colon cancer grading using infrared spectral histopathology.* Faraday discussions, 2016. **187**: p. 105-118.
31. Lasch, P. and W. Petrich, *Data acquisition and analysis in biomedical vibrational spectroscopy*, in *Biomedical Applications of Synchrotron Infrared Microspectroscopy: A Practical Approach*, D. Moss, Editor. 2010, Roal Society of Chemistry: Cambridge, UK. p. 192-225.
32. GmbH, B.O., *OPUS Spectroscopic Software: reference manual*. 2004: Ettlingen, retrieved from: http://shaker.umh.es/investigacion/OPUS_script/OPUS_5_BasePackage.pdf.
33. Naumann, D., *Infrared spectroscopy in microbiology*, in *Encyclopedia of analytical chemistry*, R.A. Meyers, Editor. 2000, John Wiley & Sons: Chichester. p. 102-131.
34. Naumann, D., *FT-IR spectroscopy of microorganisms at the Robert Koch-Institute: Experiences gained during a successful project.* Biomedical Optical Spectroscopy, 2008. **6853**.
35. Bosch, A., et al., *Fourier transform infrared spectroscopy for rapid identification of nonfermenting gram-negative bacteria isolated from sputum samples from cystic fibrosis patients.* J Clin Microbiol, 2008. **46**(8): p. 2535-46.
36. Wold, S., H. Martens, and H. Wold, *The multivariate calibration problem in chemistry solved by the PLS method*, in *Matrix pencils*. 1983, Springer. p. 286-293.
37. Breiman, L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.
38. Bassan, P., et al. *Automated high-throughput assessment of prostate biopsy tissue using infrared spectroscopic chemical imaging*. in *Medical Imaging 2014: Digital Pathology*. 2014. International Society for Optics and Photonics.
39. Großerueschkamp, F., et al., *Marker-free automated histopathological annotation of lung tumour subtypes by FTIR imaging.* Analyst, 2015. **140**(7): p. 2114-2120.
40. Menze, B.H., et al., *A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data.* BMC bioinformatics, 2009. **10**(1): p. 213.
41. Afseth, N.K. and A. Kohler, *Extended multiplicative signal correction in vibrational spectroscopy, a tutorial.* Chemometrics Intelligent Laboratory Systems, 2012. **117**: p. 92-99.
42. Kohler, A., et al., *Extended multiplicative signal correction as a tool for separation and characterization of physical and chemical information in Fourier transform infrared microscopy images of cryo-sections of beef loin.* Applied spectroscopy, 2005. **59**(6): p. 707-716.
43. Guo, S., et al., *Extended multiplicative signal correction based model transfer for raman spectroscopy in biological applications.* Analytical chemistry, 2018. **90**(16): p. 9787-9795.
44. Zimmermann, B., et al., *Characterizing aeroallergens by infrared spectroscopy of fungal spores and pollen.* PLoS One, 2015. **10**(4): p. e0124240.
45. Zimmermann, B., et al., *Analysis of allergenic pollen by FTIR microspectroscopy.* Analytical Chemistry, 2016. **88**(1): p. 803–811.
46. Zimmermann, B. and A. Kohler, *Optimizing Savitzky-Golay parameters for improving spectral resolution and quantification in infrared spectroscopy.* Appl Spectrosc, 2013. **67**(8): p. 892-902.

47. Martens, H. and M. Martens, *Multivariate analysis of quality: an introduction*. 2001: John Wiley & Sons.
48. Barker, M. and W. Rayens, *Partial least squares for discrimination.* Journal of chemometrics, 2003. **17**(3): p. 166-173.
49. Kohler, A., et al., *Interpreting several types of measurements in bioscience*, in *Biomedical Vibrational Spectroscopy* J.K. Peter Lasch, Editor. 2008, John Wiley: Hoboken, New Jersey, USA. p. 333-356.
50. Duda, R.O., P.E. Hart, and D.G. Stork, *Pattern classification*. 2012: John Wiley & Sons.
51. Preisner, O., et al., *Application of Fourier transform infrared spectroscopy and chemometrics for differentiation of Salmonella enterica serovar Enteritidis phage types.* Appl. Environ. Microbiol., 2010. **76**(11): p. 3538-3544.
52. Bağcıoğlu, M., et al., *Detection and identification of Bacillus cereus, Bacillus cytotoxicus, Bacillus thuringiensis, Bacillus mycoides and Bacillus weihenstephanensis via machine learning based FTIR Spectroscopy*. Frontiers in microbiology, 2019. **10**: p. 902.
53. Dice, L.R., *Measures of the amount of ecologic association between species.* Ecology, 1945. **26**(3): p. 297-302.
54. Sørensen, T., *A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons.* Kongelige Danske Videnskabernes Selskab, 1948. **5**(4): p. 1-34.

This paper presents a novel approach for quality testing of FTIR spectra of microorganisms based on Extended Multiplicative Signal Correction (EMSC). The approach provides better quality control compared to the standard quality test provided by OPUS from Bruker Optik GmbH. Both methods were tested on a large number of spectra of bacteria, yeasts and moulds and hyperspectral images of microorganisms where the EMSC quality test is used for image segmentation.