DOI: 10.5604/20830157.1121369

BACTERIAL PATTERN IDENTIFICATION IN NEAR-INFRARED SPECTRUM

Pavel Krepelka[1], Fernando Pérez-Rodríguez[2], Karel Bartusek[3]

[1]Brno University of Technology, Department of Theoretical and Experimental Electrical Engineering, [2]Universidad de Córdoba, Departamento de Bromatología y Tecnología de los Alimentos, [3]Academy of Sciences of the Czech Republic, Institute of Scientific Instruments

Abstract. *Microorganism identification, primary bacterial identification and pathogen detection, is important in a lot of microbial scientific areas (diagnosing of infection diseases, food protection). In this paper, the identification of the strains was performed by Near Infrared spectroscopy (wavelength from 900 nm to 2500 nm). Different techniques for classification (CVA, ANN...) were examined. It was reached to 100% accuracy on limited count of samples. Because a removing of water from sample represents a time-consuming step in sample preparation process, influence of water to spectrum was examined. Near Infrared (NIR) spectroscopy seems to be a suitable method for rapid bacteria identification. It can be used in a wide variety of food protection, medicine microbiology, bio-terrorism threats and environmental studies.*

Keywords: infrared imaging, spectroscopy, cells, absorption, near infrared spectroscopy

IDENTYFIKACJA BAKTERII W WIDMIE BLISKIEJ PODCZERWIENI

Streszczenie*. Identyfikacja mikroorganizmów, głównie identyfikacja bakterii i wykrywanie patogenów jest niezwykle istotnym zagadnieniem w wielu dziedzinach mikrobiologii takich jak: diagnozowanie infekcji czy ochrona żywności. W tym artykule dokonano identyfikacji filtracji z pomocą bliskiej podczerwieni (długość fali od 900 nm do 2500 nm). Sprawdzono różne techniki klasyfikacji (CVA, ANN ...). Osiągnięto 100% dokładność na ograniczonej liczbie próbek. Spektroskopia bliskiej podczerwieni (NIR) wydaje się być właściwą metodą do szybkiej identyfikacji bakterii. Metoda ta może być użyta do zabezpieczania żywności, w mikrobiologii, w walce przeciw bio-terroryzmowi, a także w badaniach środowiska.*

Słowa kluczowe: obrazowanie w podczerwieni, spektroskopia, komórki, absorpcja, spektroskopia bliskiej podczerwieni

Introduction

Traditional methods of bacterial identification are based on morphological examination of cells or colonies, using gram staining, examining growth characteristics and biochemical testing. These tests are time-consuming (in vitro cultivation) and subject to important sources of uncertainty. Recently, advantages of molecular techniques were exploited. Real time polymerase chain reaction is currently the most frequent molecular technique.

This highly sensitive and specific method is used to amplify DNA copies and enables its sequencing. Alternative molecular technique is infra-red spectroscopy. Advantage of infrared (especially NIR) spectroscopy lies in a rapidity and cheapness (no additional reagents or pretreatment is applied). This type of spectroscopy is a widespread technique in analytical chemistry. Recently, we can see a growing interest of the NIR spectroscopy in microbiology.

The membrane structure of bacterial cell and the ratio of lipids, proteins, and polysaccharides (IR active molecule bonds C-H, N-H, O-H) depend on the bacterial species. These changes can appear in the IR vibration spectrum. An overtone and combination bands would be detectable in the NIR spectrum. Unlike the IR spectrum, NIR spectrum is difficult to interpret, because of wide and overlapping combination and overtone peaks. To eliminate this problem, chemometrics statistical methods are utilized. Applied procedures for signal pre-processing and classification have essential influence to final identification efficiency.

1. Material and methods

In the first step, influence of water was examined. An absence of water could leads to significant acceleration of examination process. Main part of experiment was focused to verification and validation of models.

Firstly, the spectra of a solution with three different common food pathogen bacteria (*Listeria ivanovii, Escherichia coli, Salmonella*) were acquired. Bacteria have grown in tryptic soy (TSB) broth for 24 hours. Instant shaking accelerated bacteria grow. To eliminate the effect of the supernatant, the samples were centrifuged and the tryptic soy broth was changed by distilled water. This procedure removes all potential products of bacteria and provides the measurement of clear biomass. To ensure of the same level of concentration, sample absorbance of wideband (420nm-580nm) light was maintained around 0.85 (measured in 0,1 ml by Labsystems Bioscreen C). Near infra-red spectra of bacteria in an aqueous solution were acquired by diffuse reflectance integrating sphere in a region 1100 nm – 2500 nm using FT-NIR Perkin Elmer Spectrum One NTS (Shelton, CT, USA) instrument.

Transparent capsule with the test suspension was directly placed on a measurement window. This technique allows collecting transmitted and scattered light, which is useful for studying chemical and physical properties of bacteria. All accessories were cleaned by 70 % ethanol between measurements. Final spectrum was consisted from 5090 points (4 cm^{-1} spectral resolution) calculated from 128 scans.

Exact optical path length was specified by mirror adapter, which defined samples thickness to 1mm. For ensure a sufficient generality in data, 15 spectra from each bacterial strain were acquired (see Fig. 1). These data was used for development and validation of models.

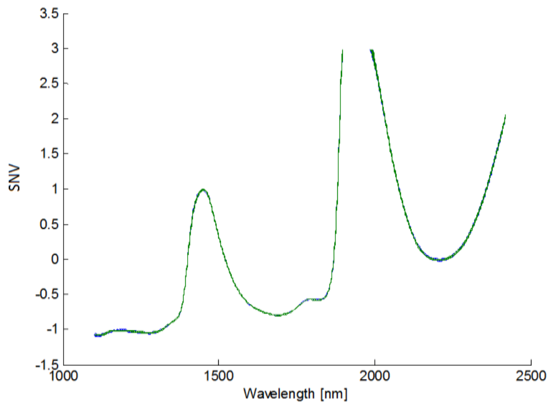


Fig. 1. Spectra of bacteria cells in water

In the second experiment, the same bacterial strains were grown TSB. Samples were centrifuged and resuspended in distilled water. Further, bacterial dispersion were filtered by glass fiber filter and dehydrated to remove absorption effect of water. The filter was putted directly to measuring window and covered by aluminum mirror. On each filter, five measurements on different place were carried out. Together, 20 spectra from each bacterial strain were acquired (see Fig. 2).

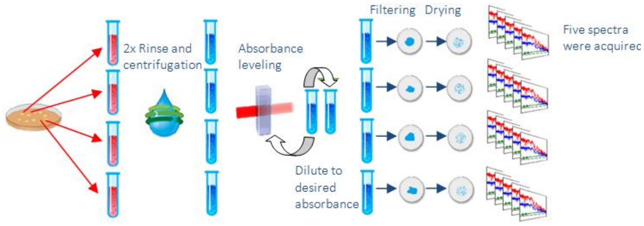

2x Rinse and centrifugation
Absorbance leveling
Dilute to desired absorbance
Filtering
Drying
Five spectra were acquirec

Fig. 2. Experimental design (figure displays procedure for one bacterial strain)

Spectra-processing techniques in a both experiments were carried out by Mathworks Matlab 8.1a. Derivation was performed by Savitky-Golay algorithm, in order to ensure smooth spectra without artifacts. For separate chemical absorption and light scattering, EMSC algorithm [5] (Extended Multiplicative Scatter Correction) was performed. EMSC algorithm can treat the NIR spectrum to be more feasible to Beer's law, which takes into consideration only chemical light absorption.

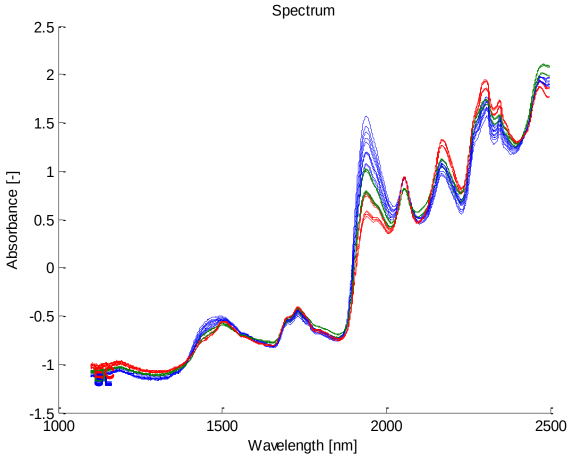


Fig. 3. Spectra of bacteria cells on glass filter. Bacteria strains are marked by different color

Because bacteria identification is classification problem, and Beer's law quantifies components, it must be realized that identification process is based on different amount of chemicals in bacterial strains. EMSC model considers wavelength-depended, multiplicative and additive effects (see eq. 1).

$$\boldsymbol{Z}_i = a_i\boldsymbol{I} + b_i\boldsymbol{m} + h_i\boldsymbol{p} + d_i\boldsymbol{\lambda} + e_i\boldsymbol{\lambda}^2 + \boldsymbol{\varepsilon}_i \qquad (1)$$

where $\boldsymbol{I}$ is a unity matrix, $\boldsymbol{m}$ is a mean spectrum (then mean of all measurements), $\boldsymbol{p}$ belongs to first canonical component and $\boldsymbol{\lambda}$ is a vector of wavelengths. The symbol $\boldsymbol{\varepsilon}$ represents a residuum caused by uncertainties measurements. This equation is minimized using the method of least squares.

Four classification methods were compared. Canonical Variate Analysis (CVA) [7] is a set of classical methods related to linear discriminant analysis. CVA is a discrimination method in which vectors maximizing the ratio of within-groups and between-groups variation.

Formalizing equation for performing a canonical correlation is relatively simple (eq. 2.). Correlation matrix (R) is formed of inverse correlation matrix (R-1) and correlation between input (independent, spectral data) and output (grouping variable) – R_{yx} R_{xy}. Space of CVA score computed in second experiment is showed in figure 4. We can clearly see separated clusters of different bacteria strains.

$$R = R_{yy}^{-1} R_{yx} R_{xx} R_{xy} \qquad (2)$$

Next often used classification procedure is Soft Independent Modeling of Class Analogies (SIMCA) [7]. In this procedure, separate model for each class is created by principal component analysis (PCA), considering only data related to one class. Each New unknown item is assessed against each of the groups depending on relative position of item to PCA systems. SIMCA was developed as alternative to principal component regression for classification purposes, like PLS-DA was developed as extension of PLS.

Partial least squares Discriminant Analysis (PLS-DA) [7] is mathematically an adequate method, because PLS is defined for quantitative variables. However, PLS-DA has been previously shown that this works well in practice. Next introduced method, Artificial Neural Network (ANN), [7] is loosely modeled brain's neural net for computation purposes.

It consists from number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs. Connections are attenuated by weights and outputs of neurons (processing element) are defined by an activation function. ANN paradigm determines parameters of ANN (num. of layers, num. of neurons, activation function, learning function) and its need to be set very carefully.

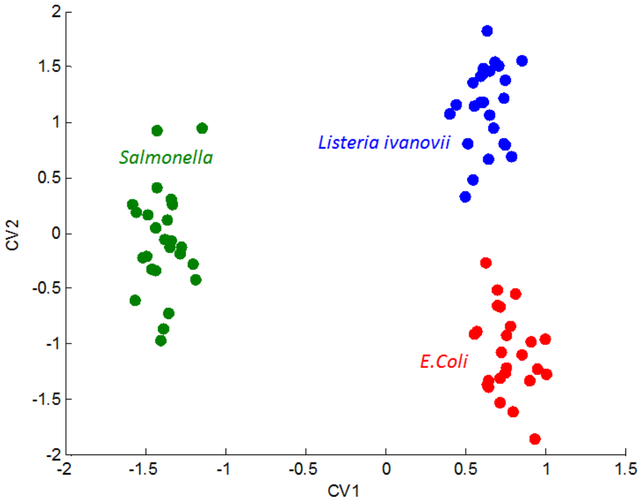


Fig. 4. Result of CVA of dehydrated samples

For increasing prediction accuracy and making the model simpler, spectrum resampling method can be applied. Spectra were divided to 509 bands. 509 models were evaluated based on the spectra without each band. By eliminating bands, model increases classification accuracy which is repeated until reaching the highest value of accuracy.

The method is similar to the well known "Jack-knifing method". An alternative me(thod is to1)select the variables, according to Analyse of variance. Only variables with a high variation between classes (dependency to variation among classes) are considered.

2. Results and conclusions

In the spectrum of aqueous solution, peaks caused of water are dominated (O-H band overtone occurred 1450 nm and wide O-H combination band at 1950 nm).

Since these peaks are significantly stronger than signal caused by bacteria cells, a detection of the signal of bacteria cells is difficult task. Despite the application of advanced statistics, all the performed models exhibit very low correct classification rates. In the Fig. 3, the pure spectra of bacteria cells are shown. High data variance according to different bacteria strains can be found at band 2300-2500 nm. It could be related to lipids and proteins contained in bacteria cells. The peaks around 1750 nm are related to C-H first overtone and it can be next indicator of proteins and their different ratios.

However, these signals are very weak in contrast of strong absorption of water. This implies that direct NIR measurements of biomass would need extreme concentrated samples to gain sufficient signal from bacteria cells without spectra saturation by water. Nevertheless, microbiology examination of aqueous specimens according to overall chemical changes is still possible.

All classification models were performed on a same set of data. Accuracy of classifiers is expressed by correct classification rate (CRR). CRR is computed by Leave-One-Label-Out Cross-Validation method. Spectra obtained from different measurements have diverse labels, totally 4 groups of measurements comprising 15 (5 for each bac. strain) spectra were used.

In every cross validation step, one group (1 measurement, i.e. 15 spectra) is excluded from training, and it is utilized for validation. Successively, all groups are selected and final CRR is computed as mean of CRR computed on all iterations.

Table 1. CRR of performed models (aqueous samples)

CRR [%]	*CVA*	*PLS-DA* (5 comp)	*SIMCA* (5 comp)	*ANN* (10x10 neur.)
Derivation + EMSC	33	36	34	36
Optimizes Der. + EMSC	35 (69x2 der. points)	35 (127 der. points)	37 (209 der. points)	35 (69 der. points)
After resampling (71 spectra points)	40	45	34	35

Classification at CVA method was accomplished by K-nearest neighbors algorithm (K=5). Technique is sensitive to the quality of input spectra, in a case of improperly chosen properties of pre-processing, algorithm exhibits lower accuracy. Whereas PLS-DA is more robust, but total accuracy (CCR) is lower. SIMCA shows better results, if EMSC is utilized.

It can be due to performing PCA which takes into account variance in spectra irrespective to classes. Strong adaptability of ANN leads to high correct classification rate even without pre-processing. For correct working, ANN needs to enough number of samples, if not this technique leads to over-fitting problem. Therefore, in case of CRR of ANN, only preliminary results are introduced.

We can see a positive effect of the resampling method. All models achieved a high score because resampling procedure is able to correct data and remove non dependent data. A similar effect was reached by optimization. The optimized model shows better results. Another losing of preprocessing methods leads to performance decrease.

Overall, results indicated a good performance of models to classify the different tested microorganisms; even though, accuracy dependences on right choose of classifier and pre-processing method.

Table 2. CRR of performed models (dehydrated samples)

CRR[%]	*CVA*	*PLS-DA* (5 comp)	*SIMCA* (5 comp)	*ANN* (10x10 neur.)
Raw spectra (SNV)	89	83	80	95
S-G 2th derivation, 69 samples	80	80	82	95
Derivation + EMSC	80	85	95	100
Optimizes Der. + EMSC	100 (69x2 der. points)	90 (127 der. points)	100 (209 der. points)	100 (69 der. points)
After resampling (71 spectra points)	100	96	100	100

Acknowledgements

This work was financed by a project of the BUT Grant Agency, No. FEKT-S-14-2545 and the EU project from the Seventh Framework Program KBBE 222738 Food, Agriculture and Fisheries, and Biotechnology and the Research Group AGR-170 HIBRO of the “Plan Andaluz de Investigación, Desarrollo e Innovación” (PAIDI).

References

[1] Alexandrakis D., Downey G., Scannell A. G. M.: Detection and identification of bacteria in an isolated system with near-infrared spectroscopy and multivariate analysis, Journal of agricultural and food chemistry, Vol. 56 (10), pp. 3431-3437.

[2] Burns, D.A., Ciurczak, E.W.: Handbook of near-infrared analysis, Third Edition, CRC Press, 2008.

[3] Cámara-Martos F., Zurera-Cosano G., Moreno-Rojas R., García-Gimeno R.M., Pérez-Rodríguez F.: Identification and Quantification of Lactic Acid Bacteria in a Water-Based Matrix with Near-Infrared Spectroscopy and Multivariate Regression Modeling, Food Analytical Methods, Vol. 5 (1), pp. 19-28.

[4] Krepelka, P.: Identification of bacteria strains via advanced methods for the statistical processing of near- infrared spectra, Proceedings of PIERS 2013, Stockholm, 2013.

[5] Naes T., Isakson T., Fearn T., Davies. T.: A user-friendly guide to multivariate calibration and classification, NIR Publications, Chichester, 2003, DOI: 10.1002/cem.815

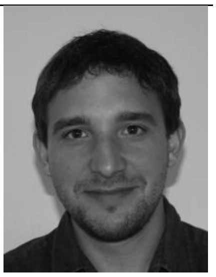
[6] Rodriguez-Saona L.E., Khambaty F.M., Fry F.S., Calvey E.M.: Rapid detection and identification of bacterial strains by Fourier transform near-infrared spectroscopy. J Agric Food Chem. 2001, Vol 49(2), pp. 574-9.

[7] Siesler H. W., Ozaki Y., Kawata S., Heise H. M.: Near-infrared spectroscopy. Principles, instruments, applications, Wiley-VCH, Weinheim, 2002.

[8] Stuard, B.: Infrared spectroscopy: fundamentals and application, Ants wiley, 2004.

[9] Thennadil S. N., Martens H., Kohler A.: Physics-Based Multiplicative Scatter Correction Approaches for Improving the Performance of Calibration Models, Appl. Spectrosc., 2006, Vol. 60, pp. 315-321.

M.Sc. Pavel Krepelka
e-mail: xkrepe01@stud.feec.vutbr.cz



P. Krepelka was born in C.Budejovice. He received his master from Electrical Engineering from the Brno University of Technology (BUT), Brno, Czech Republic, in 2011. He is currently a Ph.D student at the Dept. of theoretical electrical engineering of BUT. He made part of the research on Univ. Of Cordoba, Spain. His research interests include IR spectrometry, fibered spectroscopy, data mining, image processing and artificial intelligence. Ing. Krepelka is a member of IEEE, ICNIRS, SSJMM and IAFP. He won Deans’s prize at 2011.

Prof. Fernando Perez Rodriguez
e-mail: b42perof@uco.es

Fernando Perez Rodriguez undertook his degrees in Biological Science and in Food Science and Technology from the University of Córdoba in 1999 and 2002, respectively. He completed his PhD from the University of Córdoba (2007), which dealt with quantitative microbiological risk assessment and cross contamination in foods. He has published more than 40 articles in the most prestigious international food microbiology journals and several books concerning predictive microbiology in foods. Due to his expertise, he has participated as a scientific advisor in several expert panels at national and international level, in the Spanish Food Safety Agency and as a Risk Assessment Expert of the European Food Safety Agency (EFSA) and Food and Agriculture Organization FAO providing scientific decisions and reports. He is a member of the International Food Protection Association and the Spanish Society of Microbiology. He has also taught courses in predictive microbiology and he is professor for Food Microbiology and Hygiene at the University of Córdoba (Spain).

Prof. Karel Bartusek
e-mail: bar@ISIBrno.cz

Karel Bartusek was born in Brno, Czech Republic, in 1948. He received the Ing. (MSc.) degree in Electrical Engineering in 1973, the CSc. (PhD.) degree in 1983, DrSc. degree in 1998 and Prof. degree in 2008. He works as an independent scientific worker at the Institute of Scientific Instruments in Brno, Academy of Sciences of the Czech Republic and as a scientific worker at the Brno University of Technology, Faculty of Electrical Engineering and Communication. He is author/co-author over 200 scientific publications.

otrzymano/received: 28.09.2013 *przyjęto do druku/accepted: 12.02.2014*