Application of Multivariate Analysis, Support Vector Machines and Artificial Neural Network to the Processing of Nuclear Magnetic Resonance data of olive oil and fish oil samples for classification of geographic origin and discrimination between wild and farm fish.

ID - 121

Folgieri Raffaella[1,2], Rezzi Serge[1,3], Masoum Saeed[4], Malabat Christophe[5], Jalali-heravi Mehdi[4], Rutledge Douglas Neil[5], Axelson David E.[6], Heberger Karoly[1], Mariani Carlo[7], Reniero Fabiano[1], Guillou Claude[1]

[1]European Commission, Joint Research Centre, Institute for Health and Consumer Protection, Physical and Chemical Exposure Unit, BEVABS T.P. 740, 21020 Ispra (VA), Italy

[2]Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Milano, Italy

[3]Present address: BioAnalytical Science, Metabonomics and Biomarkers, Nestle' Research Center, P.O. Box 44, 1000 Lausanne 26, Switzerland

[4]Department of Chemistry, Sharif University of Technology, P.O. Box 11365-9516, Tehran, Iran

[5]Laboratoire de Chimie Analytique, UMR 214 INRA/INA P-G, 16, rue Claude Bernard, 75005 Paris, France

[6]MRi Consulting, 8 Wilmot St., Kingston, Ont., Canada K7L 4V1

[7]Stazione Sperimentale Olii e Grassi, Milano, Italy

Motivation

Traceability and control of origin of food products are very important for the Consumers and for the European enforcement laboratories. For instance, The high added value of olive oil makes its control an important goal for EU producers and consumers.

There is thus a need in developing analytical methods to ensure compliance with labeling, i.e.the control of geographical origin giving also support to the denominated protected origin (DPO) policy, and the determination of the genuineness of the product by the detection of eventual adulterations. Futhermore , EU regulations requires that origin, wild or farmed as well as geographic origin, of fish sold on the retail market be available to the consumers.

Modern analytical techniques such as Nuclear Magnetic Resonance (NMR) provide very informative data on the composition in fatty acids and in other constituents of vegetable oils and fish oils. The combination of 1H NMR fingerprinting with multivariate analysis provides an original approach to study the profile of these oils in relation with geographical origin of olive oil or for discrimination between wild or farm origin for fish like salmons.

Methods

Concerning the experiment on fish oil, we used Support vector machines (SVMs) as a novel learning machine in the authentication of the origin of salmon. SVMs have the advantage of relying on a well-developed theory and have already proved to be successful in a number of practical applications. The method requires a very simple sample preparation of the fish oils extracted from the white muscle of salmon samples.

Multivariate (chemometric) techniques are able to filter out the most relevant information from a spectrum, e.g. for a classification. In the experiment on olive oil samples, the principal component analysis (PCA) was carried out on the ~12,000 variables (chemical shifts) and four data sets were defined prior to PCA. Linear discriminant analysis (LDA) of the first 50 PCs was applied for classification of olive oil samples according to the geographic origin and year of production. The data analysis has been carried out with and without outliers, as well. Variable selection for LDA was achieved using: (i) the best five variables and (ii) an interactive forward stepwise manner.

Results

The use of SVMs for the discrimination between wild and farm salmon provides a new and effective method that eliminates the possibility of fraud through misrepresentation of the country of origin of salmon. The SVM has been able to distinguish correctly between the wild and farmed salmon; however ca. 5% of the country of origins were misclassified.

Using LDA on the external validation sets the correct classification of olive oil varied between 47 and 75% (random selection), and between 35 and 92% (KennardStone selection (KS)) depending on geographic origin (country) and production years.

A similar success rate could be achieved using partial least squares discriminant analysis (PLS DA). The success rate can be considerably improved by using probabilistic neural networks (PNN). Correct classification by PNN varied between 58 and 100% on the external validation sets. Other chemometric techniques, such as multiple linear regression, or generalized pair-wise correlation, did not give better results.

Email: folgieri@dico.unimi.it