3rd International Electronic Conference on Metabolomics
15-30 November 2018
chaired by Prof. Peter Meikle, Dr. Thusitha W. Rupasinghe, Prof. Susan Sumner, Dr. Katja Dettmer-Wilde
sponsored by metabolites

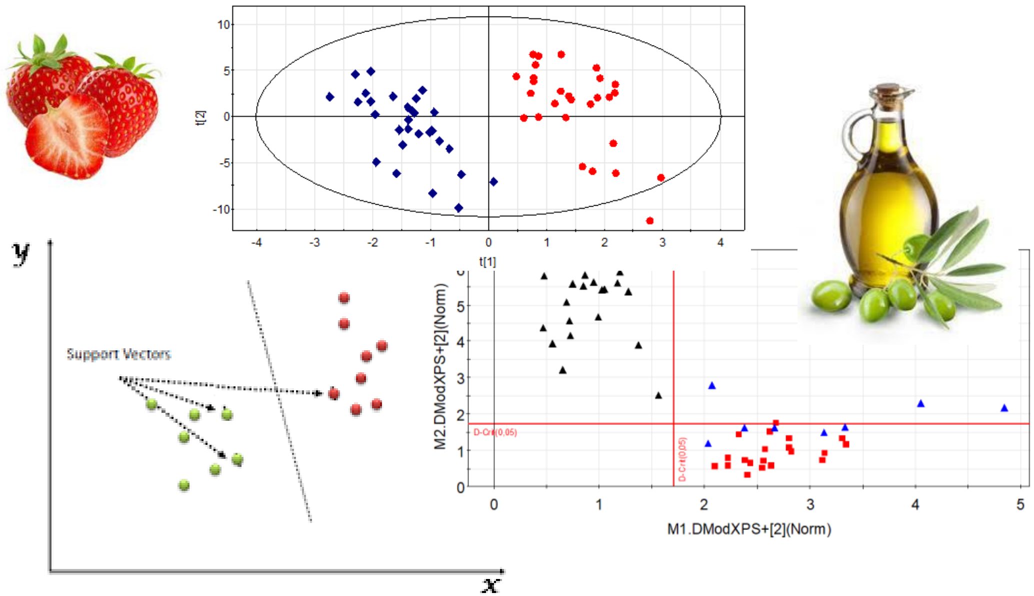
Comparison of complementary statistical analysis approaches in metabolomic food traceability

Raúl González-Domínguez [1,2*], Ana Sayago [1,2], Ángeles Fernández-Recamales [1,2]

[1] Department of Chemistry, Faculty of Experimental Sciences, University of Huelva, 21007 Huelva, Spain.
[2] International Campus of Excellence ceiA3, University of Huelva, 21007 Huelva, Spain.

* Corresponding author: raul.gonzalez@dqcm.uhu.es

Comparison of complementary statistical analysis approaches in metabolomic food traceability



Abstract:
Metabolomics generates large datasets that require the use of advanced and complementary statistical tools in order to extract the maximum amount of useful information. In this work, we show the advantages, limitations and complementarities of these techniques in food analysis, on the basis of data acquired in various traceability studies performed in our research group with strawberry and extra virgin olive oil.

Keywords: food traceability; machine learning; pattern recognition

Introduction

Omic technologies

large datasets

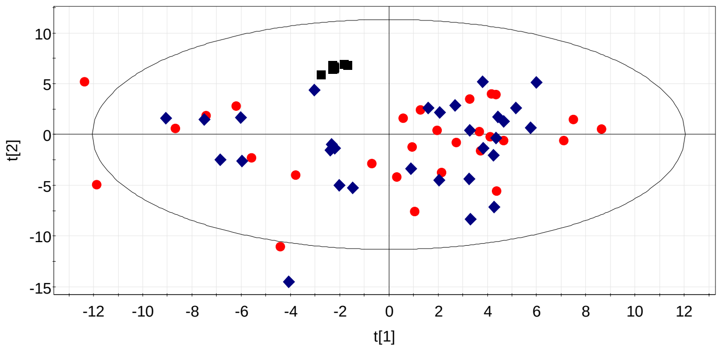
- Pattern recognition techniques: Principal component analysis (PCA), partial least squares discriminant analysis (PLS-DA), soft independent model class analogy (SIMCA)

- Machine learnig techniques: random forest (RF), support vector machines (SVM), artificial neural network (ANN)

Introduction

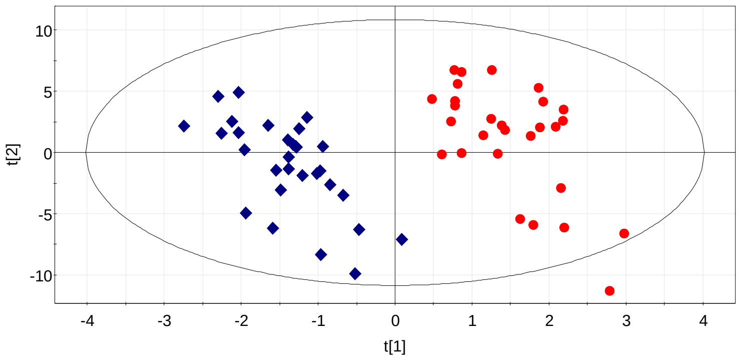
Principal component analysis

overview of data and identification of outliers and trends



Partial least square discriminant analysis

discrimination between previously defined categories

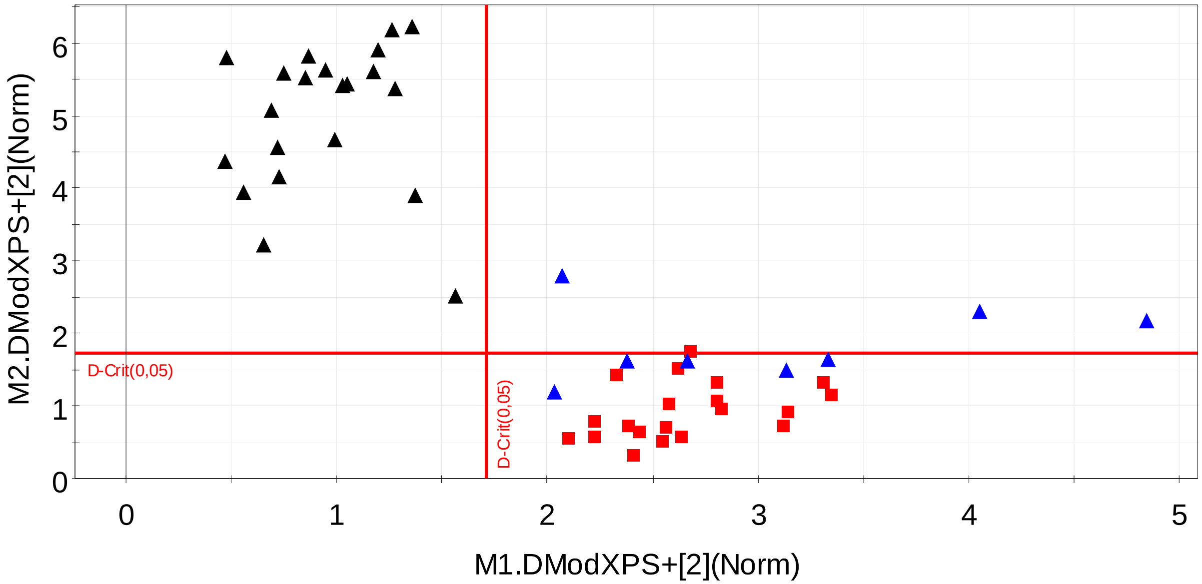


most commonly employed tools in metabolomics

Introduction

Soft independent model class analogy

Look for possible overlapping among the study groups



Introduction

Machine learning techniques

Random forest	Support vector machines	Artificial neural network

Model performance

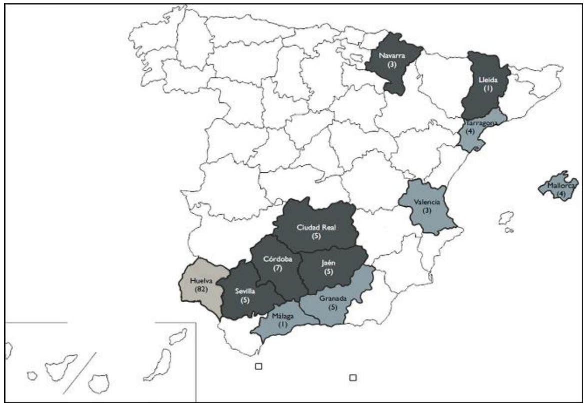
- ✓ **sensitivity** (SENS): percentage of cases belonging to a determinate class correctly classified
- ✓ **specificity** (SPEC): percentage of cases not belonging to a class and rejected by this class model

Materials and Methods

- ✓ Three varieties
- ✓ 2 macrotunnel types
- ✓ 3 conductivities of irrigation
- ✓ 3 soilless substrates

GC-MS un-targeted metabolomics [1]

LC-MS targeted metabolomics [2]



ICP-MS multielemental profiling [3]

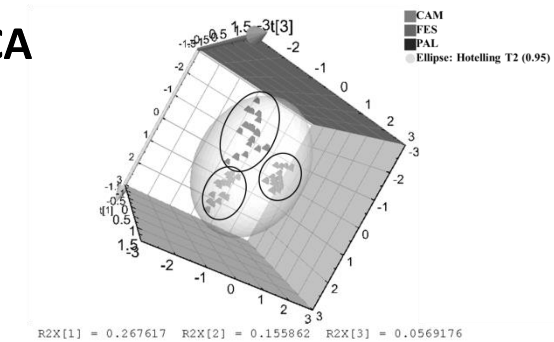
^{1}H-NMR + GC/LC profiling unsaponifiable fraction [4]

(1) Akhatou et al. Plant Physiol. Biochem. 101 (2016) 14-22
(2) Akhatou et al. J. Agric. Food Chem. 65 (2017) 9559-9567
(3) Sayago et al. Food Chem. 261 (2018) 42–50
(4) Sayago et al. Under preparation

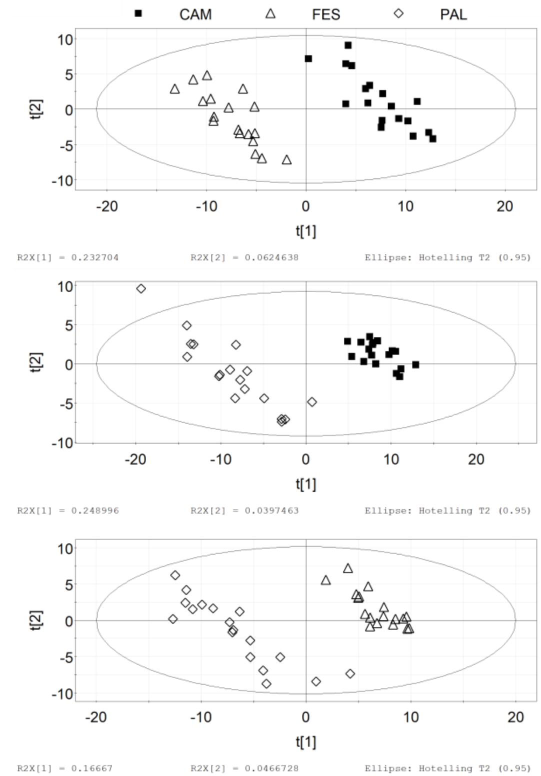
Results and Discussion

Differentiation of strawberry cultivars based on GC-MS metabolomic profiles

PCA



PLS-DA



- ✓ PCA showed good clustering of study groups
- ✓ PLS-DA to search for discriminant metabolites between varieties: sugars, organic acids, amino acids

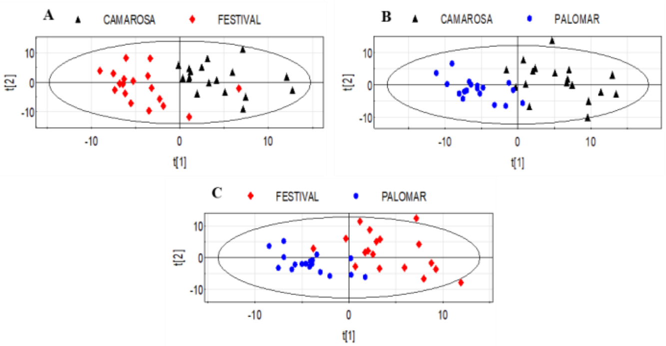
conventional statistical pipeline in metabolomics

Akhatou et al. Plant Physiol. Biochem. 101 (2016) 14-22

Results and Discussion

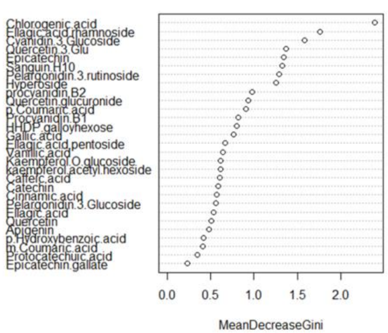
Differentiation of strawberry cultivars based on LC-MS metabolomic profiles

PLS-DA



model		'Camarosa'		'Festival'		'Palomar'		overall	
		SENS	SPEC	SENS	SPEC	SENS	SPEC	SENS	SPEC
	Cam-Fes	66.6	94.4	88.8	100			77.7	97.2
PLS-DA	Cam-Pal	72.2	100			83.3	100	77.7	100
	Fes-Pal			77.7	100	88.8	94.4	83.3	97.2
RF		100	94	94.4	100	94.4	100	96.3	96.3

RF



- ✓ Similar metabolic changes were observed in both models: anthocyanins, ellagic acid derivatives
- ✓ RF modeling provided higher sensitivity and similar specificity

Akhatou et al. J. Agric. Food Chem. 65 (2017) 9559-9567

Results and Discussion

Differentiation of olive oil provenance based on ICP-MS mineral profiles

Three predictive modelling aproaches were compared to classify EVOOs according to three geographical origins

Model	Mediterranean Coast		Inland		Huelva		Overall	
	SENS	SPEC	SENS	SPEC	SENS	SPEC	SENS	SPEC
PLS	50	100	64	98	100	100	85	98.4
SVM	77.7	100	100	94	100	100	92.7	92.7
RF	61	98	92	93.4	100	100	96.7	96.7

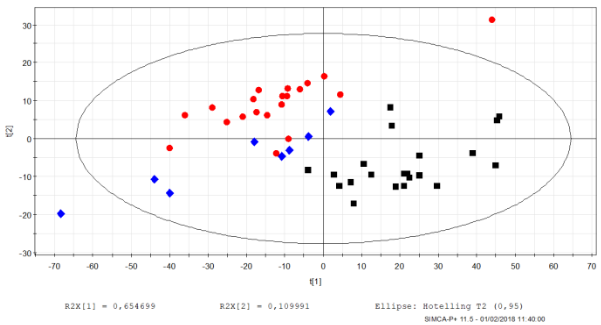
- ✓ Machine learning tools (RF and SVM) provided higher sensitivity than PLS-DA models
- ✓ Specificity was slightly higher in PLS-DA models

Sayago et al. Food Chem. 261 (2018) 42–50

Results and Discussion

Differentiation of olive oil variety based on ^{1}H-NMR and the unsaponifiable fraction

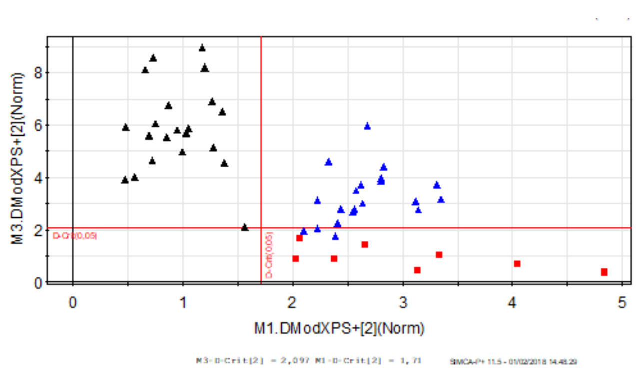
PLS-DA



■ Arbequina
● Picual
◆ Verdial

Model	Arbequina		Picual		Verdial	
	SENS	SPEC	SENS	SPEC	SENS	SPEC
SVM	100	100	100	96	87.5	100
RF	100	93.3	100	85.3	12.5	100
ANN	100	100	100	100	100	100

SIMCA



- ✓ SIMCA complements to PLS-DA with the aim of looking for possible overlapping among study groups
- ✓ Machine learning tools provide similar statistical performance

Sayago et al. Under preparation

Conclusions

- ✓ Multiple multivariate statistical tools can be complementarily employed to manage complex omic datasets
- ✓ Unsupervised PCA can be used to get an overview of data and to identify trends towards the grouping of samples
- ✓ PLS-DA is the most commonly used pattern recognition method to build classification models
- ✓ Advanced machine learning algorithms (RF, SVM, ANN) are complementary to conventional statistical techniques, which usually provide better statistical performance in terms of sensitivity and specificity