Thesis abstract. Master's degree in Agricultural Engineering (2015 June). Fernando Bueno Gutiérrez

SELECTION OF A SUBSET OF HIGHLY INFORMATIVE LOCI FOR THE ALLOCATION PROBLEM OF BOVINE ORIGIN SAMPLES TO THEIR BREEDS.

The allocation of animal origin products to their place of production is today of great interest, and specially from the year 2013 with the apparition of the “Sello de Raza” certification. Subsequently, there is a growing interest for methodologies capable of allocation tests. The simplification and cheapening of these tests will enable massive utilization. Thus, reducing frauds in the food industry. Additionally, the ability to give an added value to products based on their origin is beneficial to promote biodiversity.

Distinguishing between native Spanish breeds whose differentiation has only taken place in the last centuries, is particularly difficult and since traceability investigations began, several methods have failed in allocating individuals to their breeds. Furthermore, modern research aims to assign individuals to their breeds with as few markers as possible and this is difficult given the complexity of the internal structure of genomic data.
Genomic data, such as single nucleotide polymorphisms (SNPs) has many variables and is highly correlated. Thus, a series of analyzes are required to remove those genetic predictors whose information could invalidate the analyze for future samples.

The aim of this MSc thesis was to identify the SNPs which are more important in the differentiation of a number of cattle breeds. To achieve this, a genetic analysis was carried based on linkage disequilibrium, Hardy-Weinberg equilibrium, minor allele frequency and minimum percentage of genotyping. Subsequently, a variable selection procedure was carried to idenify the most informative predictors. For the second step, a statistical technique based on partial least squares (PLS) was considered. PLS can deal with the so-called multicollinearity problem of SNPs data. In this thesis, PLS was combined with linear discriminant analysis (LDA), which allows to rotate the data so that the differences between individuals from different groups is maximized. The methodology allows for more than 95% percentage of correct assignments to any of the eleven breeds considered using only 132 SNP markers. This demonstrates the effectiveness of PLS-LDA as a method for dimensional reduction of SNP data. Additionally, the research identified a number of chromosomal regions that may contain the most relevant loci for breeds differentiation.

Research such as this enable cheap verification of the origin of food products. Thus, contributing to higher quality of food and encouraging biodiversity.