Classification of Cattle Coat Color Based on Genotype Using Pattern Recognition Methods

M.E. Benalcázar[1,2,3], I.A. Pagnuco[1,2], D.S. Comas[1,2], P.M. Corva[4], G.J. Meschino[5], M. Brun[1], and V.L. Ballarin[1]

[1] Digital Image Processing Group, Facultad de Ingeniería, Universidad Nacional de Mar del Plata, Mar del Plata, Argentina
[2] Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, Mar del Plata, Argentina
[3] Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación, SENESCYT, Ecuador.
[4] Facultad de Ciencias Agrarias, Universidad Nacional de Mar del Plata, Balcarce, Argentina.
[5] Bioengineering Lab., Facultad de Ingeniería, Universidad Nacional de Mar del Plata, Mar del Plata, Argentina

Abstract— **Several current research projects are focused on the creation of haplotype maps to identify and describe common genetic variation in some species. Studies on haplotype maps are key in understanding how natural selection has produced genomic differences between subspecies of a given species. Important insight can be obtained by determining which variations in the genotype are associated with important phenotypical differences between individuals. Pattern recognition theory and machine learning techniques are useful tools to reveal this connection from a large amount of data provided by haplotype maps. In this work, we applied discrete classifiers and feature selection techniques for the prediction of cattle coat color from genotypes. We compared the performance of different classification rules and showed the feasibility of this approach for the prediction of phenotype based on genotype.**

Keywords— **Genotype, Single nucleotide polymorphism, Pattern recognition, Coat Color.**

I. Introduction

Several current research projects focus on the creation of haplotype maps to identify and describe common genetic variation in some species. Studies on haplotype maps are key in understanding how natural selection has produced genomic differences between subspecies of a given species.

A Single Nucleotide Polymorphism (SNP) is a DNA sequence variation occurring commonly within a specie in which a Single Nucleotide —A, T, C or G — in the genome (or other shared sequence) differs between members of biological species or paired chromosomes. Those which are located in coding sequences are likely to alter the biological function of a protein, and therefore to have an effect on the phenotype of an individual. Pattern recognition plays an important role in Genomic Signal Processing (GSP) for detection, prediction, classification, control, and statistical modeling of gene networks. One of the goals of GSP is to provide researchers with new hypothesis about biology, which can be used for systems-based applications and on confirmatory experiments, respectively [1]. Here we present an application of GSP, based on the use of pattern recognition and machine learning techniques in order to find subsets of SNPs, from a given set of SNPs, which best predicts the cattle coat color phenotype. Once identified the SNPs, they could be used in additional studies to confirm whether they are related to the underlying signaling mechanism that determines the phenotypes under study.

In this work, we want to find SNPs that may be related to coat color. Variation in coat color and spotting patterns of cattle have been extensively studied because there is evidence that animals with light-colored hair coat and darkly pigmented skin are better adapted under tropical conditions with high levels of solar radiation [2, 3].

In addition, several studies have indicated that the percentage of white on Holsteins cows may have an impact on milk yield and reproductive traits in regions of high solar radiation [4]. Any advantages of the white hair coat is due to its lowered absorption of incident solar radiation [5] resulting in reduced heat stress. Other studies reported an association between color and horn fly counts, where greater numbers of flies were observed on darker cattle [6].

The basis of coat color in cattle and all mammals is the presence or absence of melanins in the hair [7]. Coat color in cattle is largely determined by polymorphisms in the melanocortin 1 receptor MC1R gene on chromosome BTA 18 [8]. At least three alleles exist at this locus: wild type, dominant black locus, and recessive red locus. Breeds with wild-type or black allele are Holstein, Angus, Jersey, Brown Swiss; and breeds with recessive red allele are Red Angus, Charolais, Limousin, Hereford, Norwegian Red, and Santa Gertrudis [9].

Based on this goal, and previous information, we selected an initial set of 18 SNPs, or features in the language of pattern recognition, from the MC1R gene, which is suggested to be associated with the color phenotype [8]. The data for the study was obtained from the International Bovine HapMap Project [10] which contains genotype data for hundreds of individuals and thousands of SNPs. In particular, our dataset is composed of 285 samples (individuals),

132 samples from black coat individuals and the remaining 153 from read coat individuals. In this database, each feature can take 3 possible values: AA, AB, or BB, where A indicates the existence of the dominant allele and B the existence of the recessive allele, and two letters are used to represent the two copies of the chromosome.

Here, we used pattern recognition and machine learning techniques to estimate the best feature sets and their corresponding classifiers for the data available. The feature sets were ranked based on their classification error rates, limiting the number of features for each set tested to a maximum of five in order to avoid overfitting. The process of finding potential feature sets to be investigated in confirmatory experiments involves ranking such feature sets according to the performance of their optimal classifiers. Therefore, classifier design and error estimation of the designed classifiers become key factors to perform such a process successfully. From these two tasks, error estimation is perhaps the most critical task since the error estimator needs to preserve the true ordering [11-14].

II. Methods

Several classification rules have been proposed in the area of pattern recognition allowing us to design classifiers with good performance, even when there is a lack of sufficient data for posterior probability estimation [13, 15-18]. The list of rules is extensive, and we selected only some of them trading their complexity with the amount of data available for this work. This is not intended to be a complete list of possible classification approaches, but it covers many different methods well suited for 2-classes discrete classification problems of genotypes with 3 possible values for each feature. In this work the features combinations AA, AB, and BB were codified as the numbers 1, 2, and 3 respectively.

Before describing the classification rules, the notation used in this work will be introduced. Let $\boldsymbol{X}$ be a *feature vector* $\boldsymbol{X}=(x_1,...,x_d)$ with $x_i \in \mathbb{Z}, i=1,...,d$ composed of d random variables called *features*. The set of all possible feature vectors is denoted by $\mathcal{X}$. In this work, we assume $\mathcal{X}$ to be the d-dimensional space of integers $\mathbb{Z}^d$. Let Y be a binary categorical random variable with support $\mathcal{Y}=\{0,1\}$. A binary *classifier* is a function $\psi: \mathcal{X} \to \mathcal{Y}$ that assigns a label from the set $\{0,1\}$ to the feature vector $\boldsymbol{X}$; therefore, $\psi(\boldsymbol{X})$ estimates the value of Y. The set D_r represents the feature-label pairs $(\boldsymbol{X}_i, Y_i), i=1,...,n$, which contains the feature vector and its corresponding label.

A. Pyramidal Multiresolution

The pyramidal multiresolution design [19] is a combination of r classification rules, where we use a set of nested feature spaces $\{\mathcal{Z}_1,...,\mathcal{Z}_r : \mathcal{Z}_1 \supset ... \supset \mathcal{Z}_r\}$ that form a pyramid with $\mathcal{Z}_1 = \mathcal{X}$. For each feature vector $\boldsymbol{X}$, we estimate the probability of observing $\boldsymbol{X}$ by computing the relative frequency as the number of times that the feature vector has the label Y in the set D_r in the different spaces. At the application stage of a classifier designed, we label the vector $\boldsymbol{X}$ only if we have an estimate for the probability of observing $\boldsymbol{X}$ in $\mathcal{Z}_1$. Otherwise, we analyze the spaces $\{\mathcal{Z}_2,...,\mathcal{Z}_r\}$ and so on until we can label $\boldsymbol{X}$.

B. k-Nearest Neighbor (kNN)

kNN is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure. For this rule, the classier is defined for each feature vector by the label that has the highest frequency among k-nearest points (neighbors) from the training set [20].

C. Logistic Regression

This method measures the relationship between a categorical dependent variable and one or more independent variables by using probability scores as the predicted values of the dependent variable. The posterior probability distribution is modeled by using a parameterized logistic function, where the parameters are fitted by using the maximum likelihood method [15-17].

D. Linear Discriminant Analysis (LDA)

LDA is a classification technique that labels a feature vector by applying a linear transformation. Parameters of the linear transformation are estimated using the maximum likelihood method (sample-based means and covariance matrices) and the feature-label pairs [16, 17].

E. Support Vector Machines (SVM)

SVM are based on the concept of decision hyperplanes with maximal margin that define decision boundaries. A decision hyperplane partitions the feature space into two regions. The maximal-margin hyperplane, which is unique, is found by solving a quadratic optimization problem [16].

III. Proposed Method

In this work, we have a dataset $\boldsymbol{D_r}$ composed of *n* feature-label pairs $(\boldsymbol{X}_i, Y_i), i = 1,...,n$, with $n = 285$, where each vector $\boldsymbol{X}_i$ is formed by *d* features, with $d = 18$, corresponding to the eighteen SNPs selected. These SNPs are located between the base pairs 13,776,888 and 13,778,639, which corresponds to the region of chromosome 18 that contains the gene MC1R such as was defined by Stella *et. al* in [8]. The dataset used in this work comes from the *'Btau 4.0 build'* of the HapMap project [10]. Table 1 shows the identifiers, positions, and indices (for reference on this work) associated with the dataset of SNPs. This dataset contains 132 black and 153 red color samples, with proportions of 0.46 and 0.54, respectively.

In this context, the goal of this work is to find the best small subset of features (SNPs) that predicts, with high accuracy, the cattle coat color. The analysis includes the evaluation of the performance of the classifiers designed based on those features. The classification rules to be used in this work are described in Section II.

Table 1 SNPs, positions and associated indices in the dataset, corresponding to the chromosome 18 that contains the gene MC1R

SNP	Position	SNP Index
BTA-161389	14076134	1
BTA-161391	14086726	2
BTA-21794	13836376	3
BTA-21796	14047510	4
BTA-42498	13601920	5
SCAFFOLD161845_5926	13788311	6
SCAFFOLD161845_5966	13788271	7
rs29011163	13787930	8
rs29011166	13788198	9
rs29011167	13788217	10
rs29011168	13788256	11
rs29020085	13915099	12
rs29020086	13915081	13
rs29020087	13914968	14
rs29020088	13914888	15
rs29021757	13837096	16
rs29021758	13837014	17
rs29021759	13836966	18

To evaluate the performance of the designed classifiers, based on the best subset of features, we use the holdout method [15]. We split randomly the dataset $\boldsymbol{D_r}$ into 2 disjoint subsets $\boldsymbol{D_{train}}$ and $\boldsymbol{D_{test}}$, having 185 and 100 samples, respectively, maintaining the class proportions of 0.46 and 0.54. The selection of a sample size of 100 for testing ensures, in this case, a standard deviation of the estimated error upper bounded by a small value (i.e., 0.05).

Finally, the proposed method can be split into two stages: *A) Feature selection* and *B) Error estimation*. The details for each stage are described in detail in the remainder of this section.

A. Feature selection

Given the training dataset $\boldsymbol{D_{train}}$ composed of 185 samples, we need to find the best feature set for each classification rule. To avoid overfitting we limit the search for feature sets to subsets with no more than 5 features (avoiding sets with only 1 feature as well). Overfitting is related to the complexity of the model (number of parameters/features) and the number of samples available to design the classifier. High complex models will overfit small training sets and exhibit poor performance when applied to new testing sets. Because of this we are able to test for all possible combinations of only 2, 3, 4, and 5 features from the original set of 18 features, which makes a total of 12,597 features subsets to check. We rank these subsets by estimating the error of each classification rule using the *K-fold cross-validation* method with K=5 [16]. For each classification rule, the subset of SNPs with smallest error is accepted as candidate subset for classification. Note that according to the procedure described above, it is possible that we obtain different subsets of features for different classification rules.

B. Error estimation

Once we find the best subset of features for each classification rule, we use that subset to design a classifier using all the 185 samples from $\boldsymbol{D_{train}}$. The performance of that classifier is computed as its average error over the 100 left-out samples that belong to $\boldsymbol{D_{test}}$.

IV. Results

Table 2 shows the results for prediction of coat color phenotype, based on genomic data from the chromosome 18 in the positions corresponding to the gene MC1R, using the method described in Section III. For the five classification rules used (i.e., Pyramidal Multiresolution, Logistic Regression, LDA, KNN, and SVM), it displays the indices of the SNPs obtained in the feature selection stage, and estimates of the error rate, False Positive Rate (FPR), and False Negative Rate (FNR) computed on the hold-out dataset $\boldsymbol{D_{test}}$.

Analyzing the results presented in Table 2, the classification rule with best performance was LDA with an error of 21%. Among the SNPs selected as best predictors of the

coat color phenotype, the SNPs with indices 1, 5, 12, and 14 are the most frequent. These SNPs are the identifiers '*BTA-161389'*, *'BTA-42498'*, *'rs29020085'*, and *'rs29020087'*. The performances of the other classification rules are all above 74%. It should be noted that SVM is the only rule that needed only 4 SNPs to reach maximum performance.

Table 2 Classification results of the coat color phenotype of the 5 classification rules used. Error rate, False Positive Rate (FPR), and False Negative Rate (FNR) are shown.

Method	SNP indices	Error	FPR	FNR
Pyramidal Multiresolution	(5,11,12,14,18)	23%	25,92%	19,56%
Logistic	(1,3,8,12,17)	26%	33,33%	17,39%
LDA	(1,5,12,14,16)	21%	24,07%	17,39%
kNN	(1,5,12,13,14)	22%	24,07%	19,56%
SVM	(1,3,11,17)	26%	31,48%	19,57%

V. Conclusions

According to the results for the five classification rules tested here, the rule with the lowest classification error is LDA (21%). Although this is still a large error rate, it shows the feasibility of this approach, to search for biological markers for phenotypes, coat color in this case.

The SNPs identified by this approach can be useful as a guide for future biological tests, which should confirm, or not, the influence of these SNPs in the coat color phenotype.

Although the influence of the MC1R gene in the primary determination of the coat color is already verified, this work shows which SNPs, located in this gene, are more likely to be related to the variations.

It is important to note that it is biologically shown that this phenotype is also influenced by other genes [21]. Because of this, results could be improved (i.e., decrease the error rates) including SNPs from other genes involved in this phenotype. However, a larger initial set of SNPs would make harder the feature selection process and increase the potential risk of overfitting.

To future works we will apply the same methodology in other datasets. We will try to verify the biological relevance of the SNPs selected by classification rules.

Acknowledgment

M.E. Benalcázar, I.A. Pagnuco and D.S. Comas acknowledge support from Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina.

References

1. Ridder D, Ridder J, Reinders M (2013) Pattern recognition in Bioinformatics. Brief Bioinform 14:633-647
2. Finch V A, Western D (1977) Cattle colours in pastoral herds: natural selection or social preference. Ecology 58:1384
3. Finch V A, Bennetta I L, Holmesa C R (1984) Coat colour in cattle: effect on thermal balance, behaviour and growth, and relationship with coat type. J. of Agricultural Sci. 102:141-147
4. King V L, Denise S K, Armstrong D V, et al (1988) Effects of a hot climate on the performance of first lactation cows grouped by coat colour. Journal of Dairy Science 71:1093–1096
5. Stewart R E (1953) Absorption of solar radiation by the hair of cattle. Agricultural Engineering 34:235
6. Brown A H, Johnson Z B, Simpson R B, et al (1994) Relationship of horn fly to face fly infestation in beef cattle. Journal of Animal Science 72:2264-2269
7. Searle A G (1968) Comparative Genetics of Coat Colour in Mammals. Academic Press, New York
8. Stella A, Ajmone-Marsan P, Lazzari B, et al (2010) Identification of Selection Signatures in Cattle Breeds Selected for Dairy Production. Genetics 185:1451-1461
9. Olson T A (1999) Genetics of Colour Variation, in The Genetic of Cattle. CABI Publishing, The Genetic of Cattle
10. The Bovine HapMap Consortium Consortium (2009) Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. Science 324:528-532
11. Braga-Neto U M (2009) Classification and Error Estimation for Discrete Data. Current Genomics 10:446-462
12. Atashpaz-Gargari E, Sima C, Braga-Neto U, et al (2013) Relationship between the accuracy of classifier error estimation and complexity of decision boundary. Pattern Recognition 46:1315–1322
13. Dougherty E, Dalton L (2013) Scientific Knowledge is Possible with Small-Sample Classification. Bioinformatics and Systems Biology 2013:1-12
14. Esfahani M, Dougherty E (2013) Effect of Separate Sampling on Classification Accuracy. Bioinformatics 30:242-250
15. Devroye L, Györfi L, Lugosi G (1996) A Probabilistic Theory of Pattern Recognition. Springer-Verlag, Berlin Heidelberg
16. Duda R, Hart P, Stork D (2001) Pattern Classification. Wiley-Interscience
17. Murphy K (2012) Machine Learning a Probabilistic Perspective. MIT Press
18. Bishop C (2006) Pattern Recognition and Machine Learning. Springer
19. Dougherty E, Barrera J, Mozelle G, et al (2001) Multiresolution analysis for optimal binary filters. Journal of Mathematical Imaging and Vision 14:53-72
20. Rajini N H (2011) Classification of MRI brain images using k-nearest neighbor and artificial neural network, 2011 International Conference on Recent Trends in Information Technology, Chennai, India, 2011, pp 563- 568
21. Hanna L L H, Sanders J O, Riley D G, et al (2014) Identification of a major locus interacting with MC1R and modifying black coat color in an F2 Nellore-Angus population. Genetics Selection Evolution 46:1-8

Corresponding author:
Author: Diego Sebastián Comas
Institute: Facultad de Ingeniería, UNMdP
Street: Juan B. Justo 4302
City: Mar del Plata
Country: Argentina
Email: diego.comas@fi.mdp.edu.ar