Prediction and Inheritance of Phenotypes

Antonio Bahamonde, Jaime Alonso, Juan José del Coz, Jorge Díez, José Ramón Quevedo, and Oscar Luaces

Artificial Intelligence Center. University of Oviedo at Gijón, Asturias, Spain
www.aic.uniovi.es

Abstract In the search for functional relationships between genotypes and phenotypes, there are two possible findings. A phenotype may be heritable when it depends on a reduced set of genetic markers. Or it may be predictable from a wide genomic description. The distinction between these two kinds of functional relationships is very important since the computational tools used to find them are quite different. In this paper we present a general framework to deal with phenotypes and genotypes, and we study the case of the height of barley plants: a predictable phenotype whose heritability is quite reduced.

1 Introduction

There is an increasing number of applications in which genomic information (genotypes) is functionally related to observable characteristics (phenotypes). The applications include the study of risks of human diseases and the improvement in food production.

When we look for such relationship there are two possible positive outputs. First, we can find that there is a small set (usually a singleton) of genetic markers (the smallest part of genotype descriptions) whose presence implies a disease or a useful property of food, in other words, a phenotype. In these cases, the phenotype is a hereditary characteristic. The piece of genome that contains the marker, according to Mendel's Laws, is transmitted from parent organisms to their children.

On the other hand, there is a second kind of genothype/phenotype relation. The phenotypes may be predicted from a genetic description of individuals. Predictions may have different degrees. For instance, heritable phenotypes are trivially predicted. But the opposite is not always true. In fact, if predictions are established from a high number of markers, the probability of transmitting a phenotype may be low: not all relevant markers are present in children if parents do not have two copies of all these markers (homozygotes).

The existence of predictable but not necessarily inheritable phenotypes is an interesting issue that has received little attention until a few years ago. From a biological point of view these phenotypes are complex traits driven by genetic markers which may be very distant one each other in the genome [1].

The utility of such predictions has been documented in [2]. Thus, prediction is a computable way to assess the genetic risk of a disease in healthy individuals

[3]. In livestock, it has been reported [4] that artificial selection on genetic values predicted from markers could substantially increase the rate of genetic gain in animals and plants. For instance, [5] details the benefits of artificial selection in beef cattle. The authors study the use of marker-assisted selection (MAS) by using estimates (predictions) of the effects of markers on commercial crossbred performance.

In this paper, we describe the difficulties to implement a prediction system in a context of food products. So, in the next section we review how to determine a phenotype in a food environment. In these circumstances, phenotype means assessment and usually this is done by human experts whose decisions are not always repeatable. Other times phenotypes are approximate estimations of a magnitude that may have environmental (not genetic) influences. In any case, the main characteristic of these assessments is the order. In other words, the absolute values of phenotypes are irrelevant, their utility is to compare different individuals.

In the last section, we present a set of experimental results conducted to illustrate the differences between predictions and inheritances. The experiments were carried out with a dataset of barley genotypes and phenotypes publicly available. In this section we shall propose a set of measurements to assess the goodness of predictions in this genetic context.

2 Phenotypes and Genotypes

Phenotypes are quantitative expressions of any genetic characteristic or trait. As was emphasized in the Introduction, these values have only a relative meaning. They are useful only if they can be employed to order samples according to a given trait. In this section we shall review two main kinds of phenotypes of food products: those that are given by assessments rated by experts or consumers, and those that are estimations inferred from a collection of measurements of some performance related to food quality or productivity.

On the other hand, genotypes are genetic descriptions of individuals. The final aim is to find a functional relationship between genotypes and phenotypes, if such relation exists.

2.1 Phenotypes as People's Assessments

Let us start with a typical assessment problem of food products. The paper [6] describes a methodology for establishing an objective measurement of *mushroom* quality. Four experts visually evaluated 300 mushrooms and graded them into three major and eight subclasses of commercial quality. Grader consistency was also assessed by repeated classification (four repetitions) of two 100-mushroom sets. Grader repeatability ranged from 6 to 15% misclassification.

To avoid these difficulties, a number of methods have been suggested [7]. The aim is to make as *objective* as possible the assessment. For this purpose, the best option is to rest on a computable formula that uses only a set of metric

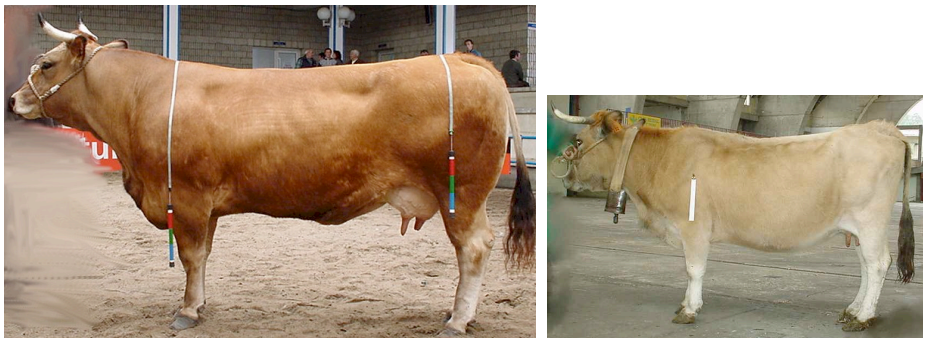

Figure 1. The assessment of the round profile of a beef cattle is a measurement of the roundness of the the curves of the buttocks. The cows in the picture are representative of an extremely high (the bigger cow) and extremely low (the smaller) round profiles

measurements. However, the complexity of the task of learning an assessment function stems from the low repeatability of human evaluations. Despite experts had been trained exhaustively and had accumulated a large and valuable body of knowledge.

Let us now recall how we dealt with the assessment of *beef cattle* as meat producers. This learning task was proposed by ASEAVA, the Association of Breeders of a beef breed of the North of Spain, *Asturiana de los Valles*. This is a specialized breed with many double-muscled individuals; their carcasses have dressing percentages over 60%, with muscle content over 75%, and with a low (8%) percentage of fat.

Even if the animals are not going to be slaughtered, the prediction of carcass value of a beef cattle is interesting since it is useful for breeders to select the progenitors of the next generation. The records of carcass value predictions can be used for the evaluation of programs of genetic selection. The growth of the scores over years of selection for specific goals can be seen as a measure of the success of the selection policy.

Traditionally, the assessment procedures were based on *visual* appreciations of well trained technicians that had to rank a number of morphological characteristics that include linear lengths of significant parts of animals' bodies. To test the reliability of these visual appreciations, we gathered (in [8]) the ASEAVA records of 2844 animals. These records include visual appreciations of the curvature of the round profile (see the curves of the buttocks in Figure 1). The ranks given by the experts of the Association of Breeders had only a poor correlation (0.70) with the EUROP ranks used to pay breeders for their carcasses. The so called *EUROP* assessment method is regulated by the European Union, and it is supposed to be correlated (for similar weights) with the curvature of the round profile.

The lack of coherence of people's assessments is not an unusual situation in beef cattle. Another application where we found assessment problems was studying *consumer tastes* about food products [9]. Consumers and experts tend to rate their preferences in a relative way, comparing objects with other samples in the same batch or tasting session. There is a kind of *batch effect* that often biases the ratings. Thus, an object presented in a batch surrounded by worse objects will probably obtain a higher rating than if it were presented together with better objects.

Therefore consumer or experts ratings cannot be interpreted as absolute assessments. They are relative values to each assessment session. From a computational point of view, this fact is the reason to reject regression methods in order to learn automatic assessment functions defined from a set of objective metric measurements of food products. From a Machine Learning perspective, if E is a set of food samples, the usefulness of people assessments (consumers or experts) is not the set of pairs sample-rating, but a set of *preference judgments* given by

$$PJ = \{(\boldsymbol{u}, \boldsymbol{v}) : \boldsymbol{u}, \boldsymbol{v} \in E, r_i^j(\boldsymbol{u}) > r_i^j(\boldsymbol{v})\}. \quad (1)$$

where $r_i^j(\boldsymbol{x})$ is the rating given by user i in session j for the product $\boldsymbol{x}$.

If phenotypes are going to be learned from people assessments, then the assessment function must be induced from datasets of preference judgments (Equation 1) using tools like those described in [10]. In [11,12] we proposed a new assessment method for beef cattle. On the other hand, in [13,14,15] we described a collection of methods to handle sensory data from consumer's opinions about food products.

2.2 Phenotypes as Approximate Measurements

There is a second kind of phenotypes, those that come from measurements of observable and quantifiable behaviors. In food products these phenotypes are related with production or performance. In the next section we are going to use a dataset of microarray descriptions of *barley plants*. The phenotypes considered are 9 traits, measured in up to 16 different environments. These traits are: α-amylase, diastatic power, heading date, plant height, lodging, malt extract, pubescent leaves, grain protein content, and yield. The number of individuals described in this dataset is quite limited, only 94 individuals.

In 8 times out of 9 the traits were represented by continuous values. The exception is the presence of *pubescent leaves*: a binary predicate. In the 8 continuos traits, to standardize the scores achieved by the learners used in these experiments, is frequent to use the percentiles of traits instead of their actual values. Therefore, all traits values range in the interval $[0, 100]$.

However, in addition to the standardization of trait measurements, the main difficulty to handle such values as phenotypes is the noise involved the estimations. This noise is due to environmental distortions. In fact, most traits, in addition to a genetic predisposition, can be increased or dismissed by factors like the season, feeding conditions, or the geological type of lands in plants.

Thus, to handle the estimations of trait measurements as phenotypes, we must have to consider such values as mere approximations. This means that no straightforward tools can be used in order to look for a computational relationship between genotypes and phenotypes.

2.3 Dealing with Genotypes

Once the phenotype of an individual i is fixed and represented by a real number (y_i), its genetic description is represented by a vector ($\boldsymbol{x}_i$) of feature values or genetic markers that are discrete values codifying the allelic composition, usually via SNP (Single Nucleotide Polymorphism). Notice that we do not consider any pedigree information.

Nowadays, microarray technology can be used to read hundreds of thousands to millions of SNPs in a single array experiment at a reasonable price. Thus, these kind of vectors are a reasonable representation of genotypes.

It is important to mention that when the genetic descriptions have physical locations in the genome sequence, the set of relevant attributes (i.e. positions in the genome sequence) somehow useful in relation to a quantitative (continuous valued) phenotype are called *QTL* (*quantitative trait loci*).

3 Prediction Tools for the Genetic Learning Task

Formally, the genetic learning task is given by a dataset

$$S = \{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}, \tag{2}$$

where $\boldsymbol{x}_i$ represents the genotype and y_i is a quantitative (i.e. a real number) phenotype of an individual i.

When phenotype values (y_i) are continuous (real) numbers, this learning task can be handled using regression tools. However, as was mentioned above, we must take into account that phenotypes are only approximate estimations. Thus, there are two possible approaches. The first identifies near values into qualitative labels. This approach rewrites S discretizing the range of phenotype values in a set of k bins of equal frequency. The aim is to redefine phenotypes in a scale of k ordered qualitative values. For instance, if we use $k = 5$, we can see this process as a translation into a ranking scale of 5 qualitative ranks: very low, low, medium, high, very high.

Rewritten in this way, S is a *classification* learning task. But in fact, it is an *ordinal regression* task since the classes are not only discrete, but also ordered. In [16], we have proposed the use of nondeterministic classifiers introduced in [8]. The aim of these classifiers is to build hypotheses that try to predict the true rank, but when the classification is uncertain the hypotheses predict an interval of ranks; a set of consecutive ranks, such that the set is as small as possible, while still containing the true rank. In symbols, a *nondeterministic hypothesis*

is a function h from the space of genotype descriptions to the set of non-empty intervals (subsets of consecutive ranks)

$$h : \mathcal{X} \longrightarrow Intervals(\{1, 2, 3, 4, 5\}). \tag{3}$$

In the case of these classifiers, it is important to register not only the proportion of classifications that include the true rank, but also the average number of ranks of the predictions. We shall limit this number to 1 or 2 in the experiments reported in the next section.

To handle a learning task S defined as in Equation (2), to clean possible *noise*, it is usual to select the most meaningful components of vectors $\boldsymbol{x}_i$ in order to establish a functional relationship with the phenotypes y_i. These components are, actually, the QTLs from a biological perspective, and they are a *selection of features* for Machine Learning nomenclature.

The feature selection methods to find QTLs are different if we have a regression (continuous phenotypes) or a classification (integer phenotypes) learning task. However, at this stage the differences are not so important. We can discretize the continuous phenotypes to obtain a classification learning task, see the preceding section. Although there are some robust methods for selecting features in regression tasks, the noise level in the genetic datasets is frequently too high to hope good results. Thus, the advantage of dealing with classification tasks is that they are implicitly taking into account that phenotype values are only approximations. Additionally, classification entails a bundle of tools to filter data sets. We shall use the filter FCBF [17] to search for a reduced number of features. This filter has been frequently used for dealing with genetic data.

Let us recall that FCBF orders the features of a learning task according to their relevance to induce class values. For this purpose FCBF uses the so called *symmetrical uncertainty* (a normalized version of the mutual information) between feature and class values, which must be discrete, not necessarily ordered. Then the filter rejects those features that are somehow redundant with other features higher in the order of relevance.

To test the goodness of a selection of features we can use the proportion of successful classifications achieved by the hypothesis learned from different subsets of features. This is an *absolute* score which is not always meaningful in the genetic case. If the phenotype is discretized in k bins, using a binary feature, it is only possible to output 2 different predictions. Thus, the maximum proportion of successful classifications is $2/k$. Taking into account this remark, let us define the *relative* accuracy of a hypothesis learned with r features on a test set $S' = \{(\boldsymbol{x'}_i, y'_i) : i = 1, \ldots, n'\}$ as follows

$$\frac{\sum_{i=1}^{n'} (y' \in h(\boldsymbol{x'}))}{n' \cdot \min\{1, \frac{2^r}{k}\}}. \tag{4}$$

Table 1. List of QTLs identified using different approaches. The column labeled by *FCBF* reports the ordered list of markers returned by the filter FCBF. The next columns report the markers selected by Composite Interval Mapping (*CIM*), single Marker Regression (*MR*), and Statistical Machine Learning (*SML*). The second part of the Table lists markers selected by CIM and MR ordered by Chromosome. In all cases, markers are represented by centimorgans (cM): distance along a chromosome

Chromosomes	FCBF	CIM	MR	SML
2H	118.9	120.7	118.9	117.8
2H	0.0		14.8	
3H	117.4	117.4	117.4	117.4
3H	74.5			
5H	65.9	68.1	71.5	68.1
4H	129.1		142.2	
1H	31.6		35.5	
7H	55.4	52.8	53.4	
1H	91.1		91.1	
5H	179.1		177.2	
3H	0.0			
7H	128.7			
6H	89.8			
2H		22.5	46.7	
2H		91.2		
4H		78.6		
4H		92.6		
6H		144.1	149.4	

4 Experimental Results

The experiments reported in this section use publicly available datasets about barley plants whose source was [18]. We used the version described in [19] that is straightforward usable by Machine Learning algorithms. The genotypes considered are a subset of a collection of more than $1,000$ RFLP, DArT and SSR markers. The genotypes of individuals of barley plants in [19] version are codified by 367 binary (0/1) features to express the presence/absence of parental allele calls. As was mentioned above, these collection of tasks report the phenotypic data for 9 traits (1 is binary, and the other 8 are continuous). The scores obtained with the 8 continuous phenotypes are described in [16]. The novelty of this paper is the discussion that follows about the genetic features relevant to predict plant heigh. All the scores reported here were obtained using a 10-fold cross validation.

Table 1 shows the list of 13 genetic markers relevant to the trait plant heigh according to FCBF and 3 other methods (see [19]). The first QTL, 2H 118.9, is also reported as relevant for this trait by the methods SML, CIM and MR [19]. This marker is inheritable, however, using this marker the quality of predictions are really very poor. The first row of Table 2 reports that only 39% of

Table 2. Scores achieved for different number of relevant features (first column). The absolute and relative proportion of successful classifications are followed by the average number of bins predicted

# features	absolute	relative	size pred.
1	0.39	0.98	1.00
2	0.50	0.63	1.23
3	0.52	0.52	1.60
4	0.65	0.65	1.57
5	0.70	0.70	1.57
6	0.69	0.69	1.54
7	0.70	0.70	1.45
8	0.71	0.71	1.44
9	0.75	0.75	1.36
10	0.68	0.68	1.22
13	0.77	0.77	1.44

Table 3. Performance of the nondeterministic classifier that predicts the rank of plant height. The rows show the proportion of successful and failed classifications of size 1 and 2 respectively

$\|h(x)\|$	success	fail	total
1	0.42	0.14	0.56
2	0.35	0.09	0.44
Total	0.77	0.23	

times it is possible to predict the true percentile interval using the value of 2H 118.9. But the relevance of this marker comes from its relative accuracy (98%). The biological interpretation is the following. With the exception of 1 sample, barleys with percentiles in $(80, 100]$ have value 0 in 2H 118.9; and all barleys with percentiles in barleys $[0, 20]$ have value 1 in 2H 118.9. In other words, this marker is a necessary but not sufficient condition to determine the heigh of a barley plant.

To reach a more useful prediction score, it is necessary to use more markers spread throughout several chromosomes (first column of Table 1). Thus the heritability of this trait is low although it may be predicted from a wide genomic description.

To illustrate the performance of nondeterministic classifiers, we included Table 3 and Figure 2. The size of predictions (first column) is the number of ranks (Equation 3) included in the intervals $h(\boldsymbol{x})$. Most of the times nondeterministic classifiers predict only one rank (56%), and in that case the proportion of successful and failed classification is 0.42 and 0.14 respectively. On the other hand, when classifiers predict 2 ranks, almost always they are right.

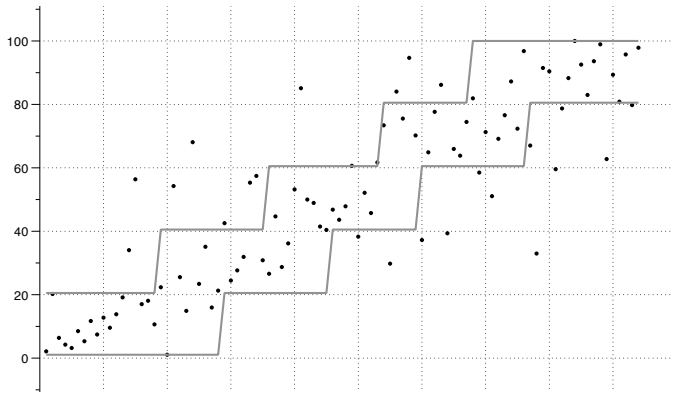

Figure 2. True plant height percentiles (•) and the intervals predicted (using a 10-fold cross validation) by a nondeterministic classifier. To ease the readability, the horizontal axis represents the indexes of barley samples ordered according to their predictions

Acknowledgements

The research reported here is supported in part under grant TIN2008-06247 from the MICINN (Ministerio de Ciencia e Innovación, of Spain). The authors also acknowledge the people who shared the barley data used in this paper, [18,19] and the software used in the experiments reported here.

References

1. Mauricio, R.: Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. Nature Reviews Genetics **2** (2001) 370–381
2. Lee, S., van der Werf, J., Hayes, B., Goddard, M., Visscher, P.: Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP Data. PLoS Genetics **4**(10) (2008)
3. Wray, N., Goddard, M., Visscher, P.: Prediction of individual genetic risk to disease from genome-wide association studies. Genome Research **17** (2007) 1520–1528
4. Meuwissen, T., Hayes, B., Goddard, M.: Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics **157** (2001) 1819–1829
5. Dekkers, J.: Prediction of response to marker-assisted and genomic selection using selection index theory. Journal of Animal Breeding and Genetics **124**(6) (2007) 331–341
6. Kusabs, N., Bollen, F., Trigg, L., Holmes, G., Inglis, S.: Objective measurement of mushroom quality. In: Proc New Zealand Institute of Agricultural Science and the New Zealand Society for Horticultural Science Annual Convention, Hawke's Bay, New Zealand (1998) 51
7. Goyache, F., Bahamonde, A., Alonso, J., López, S., del Coz, J.J., Quevedo, J., Ranilla, J., Luaces, O., Álvarez, I., Royo, L., Díez, J.: The usefulness of artificial intelligence techniques to assess subjective quality of products in the food industry. Trends in Food Science & Technology **12**(10) (2001) 370–381

8. Alonso, J., del Coz, J.J., Díez, J., Luaces, O., Bahamonde, A.: Learning to predict one or more ranks in ordinal regression tasks. Proceedings of the European Conference on Machine Learning and Practice of Knowledge Discovery in Databases (ECML/PKKD'08). Number LNAI 5211, Springer (2008) 39–54
9. Bahamonde, A., Díez, J., Quevedo, J.R., Luaces, O., del Coz, J.J.: How to learn consumer preferences from the analysis of sensory data by means of support vector machines (SVM). Trends in Food Science & Technology **18**(1) (January 2007) 20–28
10. Joachims, T.: Optimizing search engines using clickthrough data. In: Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD). (2002)
11. Bahamonde, A., Bayón, G.F., Díez, J., Quevedo, J.R., Luaces, O., del Coz, J.J., Alonso, J., Goyache, F.: Feature subset selection for learning preferences: A case study. Proceedings of the International Conference on Machine Learning (ICML'04), (2004), 49–56
12. Alonso, J., Bahamonde, A., Villa, A., Castañón, Á.R.: Morphological assessment of beef catle according to carcass value. Livestock Science **107** (2007) 265–273
13. Luaces, O., Bayón, G.F., Quevedo, J.R., Díez, J., del Coz, J.J., Bahamonde, A.: Analyzing sensory data using non-linear preference learning with feature subset selection. Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD'04), (2004) 286–297
14. del Coz, J.J., Bayón, G.F., Díez, J., Luaces, O., Bahamonde, A., Sañudo, C.: Trait selection for assessing beef meat quality using non-linear SVM. In Saul, L.K., Weiss, Y., Bottou, L., eds.: Advances in Neural Information Processing Systems 17 (NIPS '04), MIT Press (2005) 321–328
15. Díez, J., del Coz, J.J., Sañudo, C., Albertí, P., Bahamonde, A.: A kernel based method for discovering market segments in beef meat. Proceedings of the 16^{th} European Conference on Machine Learning - 9^{th} European Conference on Principles and Practice of Knowledge Discovery in Databases, (ECML/PKDD '05), (2005) 462–469
16. Luaces, O., Quevedo, J.R., Pérez-Enciso, M., Díez, J., del Coz, J.J., Bahamonde, A.: Explaining the genetic basis of complex quantitative traits through reliable prediction models. Technical report, Centro de Inteligencia Artificial. Universidad de Oviedo at Gijón (2009)
17. Yu, L., Liu, H.: Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine Learning Research **5** (2004) 1205–1224
18. Wenzl, P., Li, H., Carling, J., Zhou, M., Raman, H., Paul, E., Hearnden, P., Maier, C., Xia, L., Caig, V., Ovesná, J., Cakir, M., Poulsen, D., Wang, J., Raman, R., Smith, K., Muehlbauer, G., Chalmers, K., Kleinhofs, A., Huttner, E., Kilian, A.: A high-density consensus map of barley linking DArT markers to SSR, RFLP and STS loci and agricultural traits. BMC Genomics **7**(206) (2006)
19. Bedo, J., Wenzl, P., Kowalczyk, A., Kilian, A.: Precision-mapping and statistical validation of quantitative trait loci by machine learning. BMC Genetics **9**(35) (2008)