RanFoG: Random Forest in a java package to analyze disease resistance using genomic information

O. González-Recio[1]*and S. Forni[2]

April 5, 2010

[1]Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria. 28040 Madrid (Spain)

[2]PIC/GENUS Plc. 100 Bluegrass Commons Blvd ste 2200, Hendersonville, TN 37075 (USA)

abstract The aim of this short communication is to present a java package named RanFoG to analyze disease traits in a genomic selection context or a genomewide association study scenario. RanFoG implements random forest algorithm to analyze categorical or continuos traits. This java package has been design to determine covariate (SNP) relative importance on the phenotype expression and also to predict the outcome of yet-to-be observed records. The predictive ability performance of the program was shown using two different data sets: A simulated linear trait scenario and a real binary trait data set. Random forest presented better predictive ability than Bayes A in both type of data, and was similar to Bayesian LASSO using the simulated linear trait example. RanFoG is a developing software that is available upon request to the authors.

Keywords: Genomic selection; random forest; disease resistance; machine learning.

1 Introduction

Genomic selection offers new challenges such as the inclusion of new traits in the breeding programs. Physiological and metabolic disorders (e.g., ketosis, mastitis, metritis, scrotal hernia, Johne's disease) cause important economic loss in farms due to an increase of costs. These complex diseases hamper farms profitability and affect animal welfare. Moreover, some countries' legislation regulates antibiotic use to minimize its use in animal production due to a potential threaten of human health. Knowledge of genetic factors contributing to individual susceptibility to certain diseases will allow selection of animals genetically more resistant and even applying preventive measures to help minimizing disease development (e.g., by changing the diet, applying a determined drug, or changing management practices). This may help increasing profitability in farms.

The statistical treatment of the genetic basis of these diseases is not straightforward because most of them do not follow single-gene mendelian inheritance model, but multiple

*corresponding author: gonzalez.oscar@inia.es

genes, gene by gene interactions, and gene by environment interactions underlie most complex diseases. All these factors are seldom considered in genome-wide association studies (GWAS) or genomic selection. Limited computational resources and difficulties to capture all possible factors in a regression model are some of the restrictions to deal with. Further, phenotypes of complex diseases are generally registered in a binary (healthy/sick) or few classes shape and traditional methods are entirely based on p-value significance.

Machine learning methods is becoming more and more popular to handle these problems [1]. They aim to improve a performance measurement by repeated observation of experiences. The random forest (RF) algorithm [2] is one of the most appealing alternatives to analyze complex disease related traits using dense genomic markers information, and has been previously applied in GWAS for many human diseases. It may provide a measurement of the importance of each marker on a given disease, have good predictive performance and do not require specification of the mode of inheritance. Further, it is a fast algorithm even handling a large amount of covariates and interactions.

In the present communication we discuss the random forest approach for the inclusion of genomic information in the analyses of complex diseases, and show the use of a java program implementing this algorithm.

2 Random Forest algorithm with RanFoG

Random Forest is a massively non-parametric machine learning algorithm, robust to overfitting and able to capture complex interaction structures in the data, which may alleviate the problems of analyzing genome-wide data. We have developed a java package called RanFoG, that implements this algorithm on genomic data to predict the outcome of a given disease. RanFoG is based on a version of classification and regression trees using bootstrapped samples of the data set. Two versions are available for its implementation either on regression or classification problems, and are available upon request to the authors.

The implementation of random forest in RanFoG for genomic marker selection is described next. Let $\mathbf{y}$ (nx1) be the data vector consisting of observed records for the outcome of a given disease, and $\mathbf{X} = \{\mathbf{x}_i\}$ where $\mathbf{x}_i$ is a (px1) vector representing the genotype of each animal for p SNPs, to which T decision trees are grown. Note that main SNP effects, SNP interactions, environmental factors or combination thereof may be also included in $\mathbf{x}_i$. Each tree $h_t(\mathbf{X})$ is considered a classifier (or regressor) with the training set drawn at random with replacement from the distribution of $\mathbf{y}$ and $\mathbf{X}$. The trees are independent identically distributed random vectors, each of them casting a unit vote for the most popular outcome of the disease (or average phenotype) at a given combination of SNP genotypes. Each tree is grown as follows:

1. First, bootstrapped samples from the whole data set are drawn with replacement so that realization $(y_i, \mathbf{x}_i)$ may appear several times or not at all in the bootstrapped set $\Psi^{(t)}$, with t in $(1, T)$. This is considered as the root node of each tree.

2. Then, draw m out of p SNP markers at random, and select the SNP j, $j \in (1, ..., m)$, where $j = arg\, min_j L(\mathbf{y}, h_t(\mathbf{X}))$, i.e. SNP j is that one that minimizes a given loss function in the current node after observing its genotype. RanFoG uses entropy criterion as loss function for classification problems and the L_2 loss function for regression problems.

3. Split the node in two daughter nodes according to SNP j genotype that one individual may or may not have (e.g. individuals with the risk allele will pass to a child node, and the remaining animals will pass to the other child node).

4. Repeat steps 2-3 until a minimum node size is reached (usually <5). The predicted value of new point is the majority vote for the disease outcome at the ending nodes (for regression problems, it is the average phenotype of the individuals in the node).

5. Repeat steps 1-5 a large enough number of times to grow a random forest.

Final predictions can be made by averaging the value predicted at each tree to obtain a probability of being susceptible. In addition, RF may calculate the relative importance of each SNP, covariate or interaction. The variable importance (VI) is estimated as follows. After each tree is constructed, the out of bag samples (OOB), which are those observations left out in the bootstrapped sampling, are passed down the tree and the prediction accuracy of disease outcome is calculated using the chosen criterion (e.g. misclassification rate, L_2 loss function). Then, genotypes for the pth SNP are permutated in the OOB, and the accuracy for the permutated SNP is again calculated. The relative importance is calculated as the difference between these prediction accuracies (that from the original OOB and that of the OOB with the permutated variable). This step is repeated for each covariate and decrease of accuracy is averaged over all trees in the random forest. The variable importance provides an insight of the SNP association level with the disease. The SNPs with higher VI may be of interest for prediction of disease resistance at low marker density, candidate gene studies or gene expression studies.

This RF design was implemented on two data examples (classification and regression problems) using the java package RanFoG:

Example 1. Regression. RCS simulated data

QMSim software [3] was used to simulate a reference population (40,195 animals) and a testing population (1005 animals which were progeny of the reference population) for a heritability trait of 0.25. The reference and the validation population were genotyped using 9990 markers. Parameters of the simulations may be found in Jimenez-Montero et al. [4]. Then, 2500 individuals from the reference population were randomly extracted and used in the analyses.

Example 2. Classification. Boars scrotal hernia classification in a real data set

Data were provided by PIC North America, a Genus Plc company. The data set contained records of scrotal hernia (SH) incidence (scored as 0 or 1) in 986 animals from a commercial line born in elite genetic nucleus, where environmental conditions were controlled and risk of infections is low. Genotypes of all animals with phenotypic records were obtained for 6742 SNPs located in different genomic regions including those identified as candidate regions in previous research. After genotype editing following Ziegler et al. [5], 5302 SNPs were retained and 923 animals were used. For each individual and main effect for SNP jth, we defined two covariates x_j^1 and x_j^2, with $x_j^1 = 1$ if the genotype was aa (0, otherwise), and $x_j^2 = 1$ if the genotype was AA (0, otherwise). Analysis was performed in a cross validation scenario leaving 15% of youngest animals out as testing set. Estimates reported are the average of twenty five independent RF runs.

Bayes A [6] was used as benchmark. A threshold version of Bayes A (TBA) was used on the SH data. Additionally, Bayesian LASSO [7] was also used to analyze the

simulated RCS data. Predictive ability of methods was based on predictive accuracy in the testing set measured through mean squared error (MSE) and correlation between predicted genomic value and true genomic value (TGV), rTI. Sensitivity and specificity were also used as predictive ability parameter in the SH data. These parameters provide information about performance of methods to correctly predict resistant and susceptible animals. In the SH database, the true genomic values were unknown. Hence, the EBVs from the PIC routine genetic evaluation were assumed to be the TGV. These routine genetic evaluations are implemented with BLUP and data on millions of animals were used. The pedigree and phenotypic data included records collected over 15 years in more than 20 countries. EBV accuracies of genotyped animals ranged between 0.50 and 0.98. We agree with criticism about the realism of this assumption under the presence of important non-additive genetic effects and for low accurate EBV. However, this is the closest value to TBV that, at this moment, we can obtain. This approach may reflect the benefits of genomic selection to the industry as it usually utilizes these EBV as TBV despite of low accuracy. To minimize the issue of this approximation, true breeding values were classified as susceptible or non susceptible following the categorical nature of the phenotype. Non susceptible animals were those in the lower α percentile of the EBV distribution, whereas those in the upper $1-\alpha$ percentile were considered as susceptible ($\alpha \epsilon \{5, 10, 25, 50\}$). Lower values of α selected the more extreme animals, and smaller approximation error is expected.

3 Results

Table 1 shows the Pearson correlation and MSE between predicted genomic value and TGV from each model on the simulated RCS data. The performance of Random Forest compared favourably with Bayes A and was similar to Bayesian LASSO. Both Bayesian LASSO and RF showed larger Pearson correlation (0.57 and 0.59, respectively) and lower MSE (0.15 and 0.19, respectively) than Bayes A ($r = 0.33; MSE = 0.44$). SNP estimates in the higher 99th percentile were closer to the true QTLs using Bayesian LASSO. The VI estimates in the 99th-percentile from RF detected a lower number of SNPs than Bayes A and Bayessian LASSO, clustering around QTL's with larger effect (Figure 1).

Table 2 shows results obtained with the real SH data. Both methods (TBA and RF) were more accurate in correctly detecting the most extreme animals, i.e. lower misclassification rate, and larger r_ϕ, sensitivity and specificity were achieved at lower values of α. RF achieved misclassification=0, r_ϕ=1, sensitivity=1 and specificity=1 at α=5, which means a perfect classification of most extreme animals. At this α level TBA showed misclassification rate=17%, r_ϕ=0.71 and was less sensitive than RF. Random Forest achieved up to a 81% larger r_ϕ than TBA at α=25. The machine learning algorithm was more specific but less sensitive than TBA. Therefore, detection of susceptible animals was done more accurately using RF, whereas TBA detected resistant animals in a better manner. Nonetheless, the choice of one or other method to correctly detecting susceptible or resistant animals may depend upon the problem. Results showed that both RF and TBA may misclassify intermediate animals. However, in a disease resistance genomic selection context we are mainly interested in correctly detecting the most susceptible or resistant animals (lower α values), and RF seemed to perform slightly better than TBA to detect susceptibility to SH in this population.

Table 1. Pearson correlation and mean squared error (MSE) between predicted and true genomic values using Bayes A, Bayesian LASSO (BL) and Random Forest (RF) on the RCS simulated data.

	Bayes A	BL	RF
Pearson correlation	0.33	0.57	0.59
MSE	0.44	0.15	0.19

Table 2. Sensitivity, specificity, Phi correlation and misclassification rate of predictions using a threshold version of Bayes A (TBA) or Random Forest (RF) on the extreme animals in the EBV α and $(1-\alpha)$ percentile of the real scrotal hernia data ($\alpha \epsilon \{5, 10, 25, 50\}$).

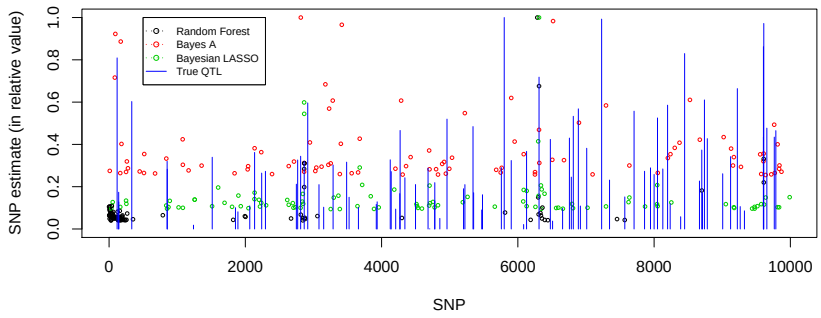
Parameter	Method	α (number of records)			
		5 (12)	10 (79)	25 (98)	50 (138)
Specificity	RF	1	0.88	0.78	0.79
	TBA	1	0.71	0.58	0.58
Sensitivity	RF	1	0.52	0.52	0.46
	TBA	0.75	0.58	0.58	0.56
Phi correlation	RF	1	0.33	0.29	0.26
	TBA	0.71	0.24	0.16	0.13
Misclassification rate	RF	0	41	39	38
	TBA	17	39	42	43

4 Remarks

Genomic selection and GWAS on disease resistance in livestock species are likely to become even more important in the next years because of worldwide guidelines on animal and human health. In general, RF seems an elegant method with an interesting predictive ability for disease resistance studies using whole genome information. In general, RF outperformed the classification predictive ability of TBA, and was better than Bayes A for regression problems with similar performance to Bayesian LASSO. Moreover, RF is a fast algorithm and robust to overfitting. These features make RF an appealing method to evaluate disease resistance in a genomic selection context. Although, further research on the method is necessary. Sires or individuals may be genomically evaluated for disease susceptibility, selecting those genetically more resistant to complex diseases affecting farm profitability. Random Forest, as well as other methods, will need to be modified and adapted to deal with some of these issues and challenges at incorporating genomic information in the analyses of disease traits.

Finally, strategies to include genomic selection for disease resistance on breeding programs have to be studied, developed and implemented.

Figure 1. Relative effect estimate and position of SNPs in the higher 99th-percentile estimates from each model, and position and relative magnitude of the true QTLs.



Acknowledgements

The authors wish to thank PIC/Genus Plc. for providing the scrotal hernia data and J.A. Jiménez-Montero for sharing the simulated data.

References

[1] Szymczak S., J.M.Biernacka, H.J.Cordell, O.González-Recio, I.R.König, H.Zhang, and Y.V.Sun. Machine learning in genome-wide association studies. *Genetic Epidemiology*, 33(S1):S51–S57, 2009.

[2] Breiman L. Random forest. *Machine Learning*, 45(1):5–32, 2001.

[3] Sargolzaei M. and F.S. Schenkel. Qmsim: a large-scale genome simulator for livestock. *Bioinformatics*, 25:680–681, 2009.

[4] Jiménez-Montero J.A., O. Gonzalez-Recio, and R. Alenda. Genotyping strategies for genomic selection in dairy cattle. *XV Reunión Nacional de Mejora Genética Animal. Vigo*, 2010.

[5] Ziegler A., I.R. Konik, and J.R. Thompson. Biostatistical aspects of genome-wide association studies. *Biometrical Journal*, 50:1–21, 2008.

[6] Meuwissen T. H. E., B.J. Hayes, and M.E Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157:1819–1829, 2001.

[7] Park T. and G. Casella. The bayesian lasso. *Journal of American Statistical Association*, 103:681–686, 2008.