ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics

SHORT COMMUNICATION

doi: 10.1111/age.12208

Locus minimization in breed prediction using artificial neural network approach

M. A. Iquebal[1a], M. S. Ansari[2a], Sarika[1a], S. P. Dixit[3], N. K. Verma[4], R. A. K. Aggarwal[5], S. Jayakumar[4], A. Rai[1] and D. Kumar[1]

[1]Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, Library Avenue, PUSA, New Delhi 110012, India. [2]Jamia Millia Islamia, Jamia Nagar, New Delhi 110025 India. [3]Animal Biotechnology Division, National Bureau of Animal Genetic Resources, Karnal Haryana 132 001, India. [4]Animal Genetics Division, National Bureau of Animal Genetic Resources, Karnal Haryana 132 001, India. [5]Animal Genetic Resources Division, National Bureau of Animal Genetic Resources, Karnal Haryana 132 001, India.

Summary

Molecular markers, *viz*. microsatellites and single nucleotide polymorphisms, have revolutionized breed identification through the use of small samples of biological tissue or germplasm, such as blood, carcass samples, embryos, ova and semen, that show no evident phenotype. Classical tools of molecular data analysis for breed identification have limitations, such as the unavailability of referral breed data, causing increased cost of collection each time, compromised computational accuracy and complexity of the methodology used. We report here the successful use of an artificial neural network (ANN) in background to decrease the cost of genotyping by locus minimization. The webserver is freely accessible (http://nabg.iasri.res.in/bisgoat) to the research community. We demonstrate that the machine learning (ANN) approach for breed identification is capable of multifold advantages such as locus minimization, leading to a drastic reduction in cost, and web availability of reference breed data, alleviating the need for repeated genotyping each time one investigates the identity of an unknown breed. To develop this model web implementation based on ANN, we used 51 850 samples of allelic data of microsatellite-marker-based DNA fingerprinting on 25 loci covering 22 registered goat breeds of India for training. Minimizing loci to up to nine loci through the use of a multilayer perceptron model, we achieved 96.63% training accuracy. This server can be an indispensable tool for identification of existing breeds and new synthetic commercial breeds, leading to protection of intellectual property in case of sovereignty and bio-piracy disputes. This server can be widely used as a model for cost reduction by locus minimization for various other flora and fauna in terms of variety, breed and/or line identification, especially in conservation and improvement programs.

Keywords breed assignment, DNA markers, goat breed, webserver

For successful conservation and long-term sustainability of existing breeds or extant populations, identification of pure breeds is imperative. When two breeds resemble each other phenotypically, identification of the breed becomes subjective. Poor reproductive performance with a high mortality rate is evident when native goats are cross-bred with exotic goat breeds, *viz.* Alpine, Saanen and Boer (Rai *et al.* 2005), necessitating selective breeding of the true-to-breed-type animals and making breed identification using molecular tools a critical requirement. In cases when the degree of breed admixture is not conspicuously visible, it is hard to differentiate between true-to-breed type and an 'admixtured breed'. In such situations, microsatellites and SNPs are strong molecular markers for breed identification. Molecular-marker-based identification of breed is possible even from small amounts of biological tissue or germplasm, such as ova and semen, that show no evident phenotype. Breed identification through markers is one of the promising means of establishing genetic identity, kinship of an animal

Address for correspondence

D. Kumar, Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, Library Avenue, PUSA, New Delhi 110012, India.
E-mail: dineshkumarbhu@gmail.com

[a]These authors contributed equally.

Accepted for publication 8 July 2014

and product traceability (Dalvit *et al.* 2007). Microsatellite-DNA-marker-based breed assignments has been reported in various domestic animals such as cattle (Blott *et al.* 1999; Maudet *et al.* 2002), sheep (Arranz *et al.* 2001; Niu *et al.* 2011), goat (Serrano *et al.* 2009; Hoda *et al.* 2011), pig (Fan *et al.* 2005), horse (Bjornstad & Roed 2001), dog (Toskinen & Bredbadka 1999), poultry and rabbit (Gotz & Thaller 1998). Although recent server-based reported breed prediction (Iquebal *et al.* 2013) has reduced the cost of reference data through its online availability, the method described, especially for less differentiated populations, needs a larger number of loci. We report here the successful use of an ANN approach, which is a nonparametric technique, along with the server for further drastic cost reductions by locus minimization without compromising the accuracy of identification.

In general, a simple ANN model consists of three layers of nodes, *viz.* input, hidden and output layers. It allows for the connection of each node in one layer with every other node in the next layer. Training of the network is accomplished using several algorithms that estimate the functional relationship between inputs and outputs using supervised learning and by means of estimating the weights associated between the nodes at all iterations to minimize the sum of squared errors. Important activation functions, such as identity, tanh, logistic, exponential and sine, are used. The most popular form of ANN architecture is the multilayer perceptron (MLP), which is a generalization of the single-layer perceptron. MLP is a feed-forward neural network architecture with unidirectional full connections between successive layers. It consists of a set of source nodes that constitute the input layer, one or more hidden layers of computation nodes and an output layer. The input signal propagates through the network in a forward direction on a layer-by-layer basis. Further, in general, learning is used to describe the process of finding values of weights. A learning algorithm adjusts connection weights until the system converges to approximately reproduce the output. Optimal weights may be obtained by using a gradient descent algorithm (GDA), the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm or a conjugate gradient descent algorithm (CGDA) with a view toward minimizing the sum of the squared error function of network output (Hassoun 2003; Yegnanarayama 2006).

To achieve locus minimization using ANN methodology, we used 51 850 samples of allelic data of microsatellite-marker-based DNA fingerprinting on 25 loci covering 22 registered goat breeds of India. We report locus minimization, using this novel computational approach, which was not possible using earlier reported methods (Dixit *et al.* 2012; Iquebal *et al.* 2013).

From 22 goat breeds, that is, Blackbengal, Ganjam, Gohilwari, Jharkhandblack, Attapaddy, Changthangi, Kutchi, Mehsana, Sirohi, Malabari, Jamunapari, Jhakarana, Surti, Gaddi, Marwari, Barbari, Beetal, Kanniadu, Sangamnari, Osmanabadi, Zalawari and Chegu, selected from diverse geographical regions, a total of 1037 blood samples were collected from genetically unrelated animals. The number of samples from each breed was 48, except for Kutchi, Chegu and Zalawari, where samples numbered 46, 43 and 36 respectively. From each animal, 5–6 ml of blood were obtained by jugular venipuncture using vacuum tubes treated with 1.5% ethylene diamine tetra acetic acid (EDTA) as an anticoagulant. Genomic DNA was isolated as per the method described by Sambrook *et al.* (1989) with minor modifications whereby isolated DNA pellets were resuspended in 150 ml of TE buffer (10 mM Tris-HCl, pH 8.0; 1 mM of EDTA and 50 mg/ml of RNAse). After checking the quality and quantity of the DNA, it was diluted to a final concentration of 50 ng μl/1 in water and stored at 4 °C.

Twenty-five primers were used for data generation, *viz.* ILST008, ILSTS059, ETH225, ILSTS044, ILSTS002, OarFCB304, OarFCB48, OarHH64, OarJMP29, ILSTS005, ILSTS019, OMHC1, ILSTS087, ILSTS30, ILSTS34, ILSTS033, ILSTS049, ILSTS065, ILSTS058, ILSTS029, RM088, ILSTS022, OarAE129, ILSTS082 and RM4. Only forward primers at the 5′ end of each pair were labeled with one of the four fluorophores, that is, FAM (Blue), VIC (Green), NED (Yellow) and PET (red).

For allele data generation, polymerase chain reaction (PCR) was carried out on 50 ng of genomic DNA in a 25-μl reaction volume. The reaction mixture consisted of 200 μM of each dNTP, 50 nM of KCL, 10 mM of Tris-HCL (pH 9.0), 0.1% Triton X-100, 2.0 mM of $MgCl_2$, 0.75 unit of Taq DNA polymerase and 4 ng/μl of each primer, and a PTC-200 PCR machine (MJ Research) was used. The 'touchdown' PCR protocol used was an initial denaturation of 95 °C for 3 min, three cycles of 95 °C for 45 s and 60 °C for 1 min, three cycles of 95 °C for 45 s and 57 °C for 1 min, three cycles of 95 °C for 45 s and 54 °C for 1 min and 20 cycles of 95 °C for 45 s and 51 °C for 1 min with a final extension at 72 °C for 5 min. PCR products were checked for amplification in 2% agarose gel by electrophoresis followed by UV light visualization in ethidium bromide staining. After determining the optimal pooling ratio and dilution ratio for a set of primers, the PCR products were mixed in a ratio of 1:1.5:2:2 of FAM (blue), VIC (green), NED (yellow) and PET (red) labels respectively. Further, 0.5 μl of this mixture was combined with 0.3 μl of Liz 500 as an internal lane standard (Applied Biosystems) size calibrator and 9.20 μl of Hi-Di Formamide per sample. This mixture was denatured at 95 °C for 5 min. All denatured samples were run on an ABI 3100 Avant automated DNA sequencer (Applied Biosystems). Microsatellite allele size data were generated by GENEMAPPER software (version 3.0; Applied Biosystems) using electropherograms, which were drawn by the GENESCAN software of the automated DNA sequencer (Kumar *et al.* 2009).

It was observed that, for the Indian goat dataset, the 25 loci were imperative to reach 99% accuracy by the Bayesian

method using GENECLASS2 (Piry *et al.* 2004). The mean number of alleles and effective number of alleles were obtained using POPGENE (Yeh *et al.* 1999), and F_{ST} values for each locus were computed using FSTAT (Goudet 2002). Henceforth, the ANN approach was applied to build an accurate model for the classification of goat breeds. Both types of network – MLP and radial basis function (RBF) – were trained using 51 850 samples of microsatellite data on 25 loci to develop the best model for breed identification. The training was performed using STATISTICA ver. 6. Before training, available data from 1037 observations were divided into two subsets randomly: (i) a training set comprising 831 (80%) observations, to be used for computing and updating the network weight and biases, and (ii) a test set comprising the remaining 206 (20%) observations. The MLP and RBF networks were trained using GDA, BGFS algorithm and CGDA with a view to minimize entropy as an error function of the network output. Several learning rate values for training the networks and activation functions, *viz.* identity, tanh, logistic, exponential and sine, for hidden units and output units were tried. A feature selection technique was applied to minimize the number of loci to decrease the genotyping cost. Fivefold cross-validation was implemented to estimate the error rate.

The prediction quality of the model was examined through evaluation measures such as accuracy, Mathew's correlation coefficient (MCC), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) (Hamada *et al.* 2010). The values were estimated by constructing a 22×22 confusion matrix (contingency table). The server was developed using ASP.NET with C# code at the back end, which was generated using STATISTICA. HTML. Java scripts were used and this was implemented as a webserver.

Table 1 Training and testing performances of various models.

Models	Training performance	Testing performance	Hidden activation	Output activation
MLP 355-18-22	96.63	94.17	Tanh	Softmax
MLP 355-19-22	97.25	78.06	Tanh	Logistic
RBF 355-22-22	93.64	64.94	Gaussian	Softmax
RBF 355-18-22	92.15	61.26	Gaussian	Softmax
MLP 355-17-22	95.21	88.06	Tanh	Softmax
MLP 355-15-22	87.21	76.13	Identity	Tanh

We used the well-established statistical tool GENECLASS2 for breed prediction. We found the maximum accuracy of 99% using the Bayesian approach, but locus minimization could not be achieved due to poor F_{ST} values of the loci. Using the same dataset, we found the ANN approach to be much superior in terms of locus minimization up to nine loci.

Artificial neural network methodology was applied very successfully over the 51 850 allelic data points of Indian goat breeds. It was observed that the MLP neural network (MLP 355-18-22) outperformed other methods (Table 1) with fivefold cross-validation. The training and testing performance reported were 96.63% and 94.17% respectively with the BFGS training algorithm, Tanh hidden activation function and Softmax output activation function. The overall performance of the model was 96.14%. Of the several tried learning rate values, the best result was obtained at 0.1. The computed average values for sensitivity, specificity, PPV, NPV, accuracy and MCC were 96.2%, 99.8%, 96.3% and 99.8%, 99.6% and 96.0% respectively. The numbers of loci were optimized to nine from 25 (Fig. 1). There are even cases of domestic animal breed predictions using three loci in horse (Bjornstad & Roed 2002). A minimum number of loci with high accuracy is always desirable, and such success comes when loci are highly differentiated, that is, high F_{ST} values; for example, in the case of horse, F_{ST} is 0.2–0.25. The maximum individual assignment success with F_{ST} of 0.18 across 10 loci has been reported in dog (Koskinen 2003). In our case, the F_{ST} values for all the 25 loci were observed to lie between 0.049 and 0.394 with overall genetic differentiation of 16.48%. Our reported nine loci were re-evaluated by the Bayesian method using GENECLASS2. It was found that nine breeds

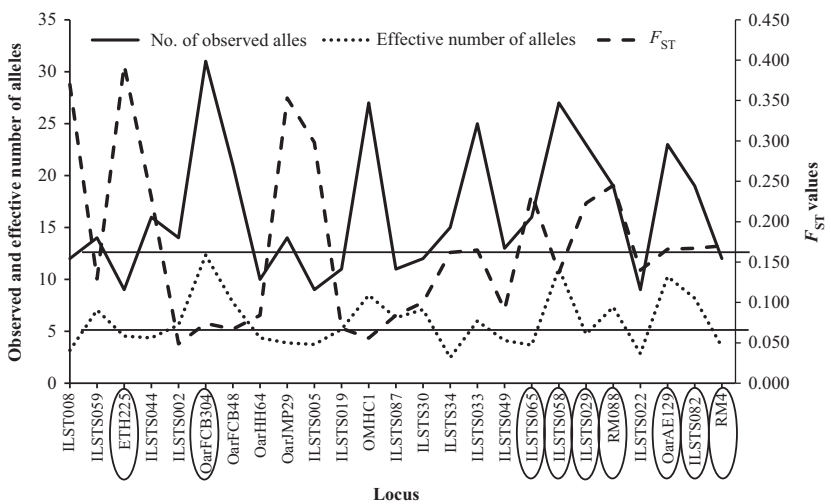

Figure 1 F_{ST}, mean number of alleles and effective number of alleles values based on 25 loci of Indian goat breeds (circled loci were selected for breed prediction).

showed <90% accuracy, some of them as low as 42%. Applying ANN methodology clearly resolves the accuracy issue with a minimum number of markers. Further, the mean number of alleles and effective number of alleles were obtained using POPGENE (Fig. 1).

An interesting observation was found between F_{ST} and accuracy. We found the best signature ability in nine loci (represented by the circle in Fig. 1) having higher F_{ST} values (>0.1) with one exception, that is, 29 *OarFCB304*, which had a F_{ST} of <0.1. This one locus having a lower F_{ST} still contributed to breed 'signature ability' due to informativeness compensation by a higher number of observed and effective number of alleles. Contrary to this, we found that, of 25, only one locus, that is, *OarJMP29*, had a relatively high F_{ST} value (>0.1) but did not contribute to the signature ability of breeds. F_{ST} is relative loci differentiation, and when two populations or breeds have exactly same allelic frequency, then F_{ST} is zero. When all loci in both breeds (populations) have unique private alleles, then the F_{ST} equals one, which is just a theoretical situation (Weir & Cockerham 1984). In reality, the value is always between zero and one for every locus. In our case, only one locus, *OarJMP29* (with higher F_{ST}), was not able to differentiate, which is a situation specific for our dataset only. Such an observation is very much expected, especially when the number of breeds in the data panel is on the higher side (we were comparing 22 breeds/population). Moreover, the DNA signature of breeds is a 'statistical signature' based on allele type, relative frequency and its relative distribution; thus, one or two such loci will not 'dilute' the signature ability of the set of loci finally selected (MacHugh *et al.* 1998; Bjornstad & Roed 2002; and Koskinen 2003).

Our analysis also adds a new dimension, such that in the rare case when high F_{ST} is not supported by a higher effective number of alleles, then the breed signature-making ability of locus is lost. We found that both F_{ST} and the effective number of alleles have threshold value (as shown in Fig. 1) to be fit for 'signature loci' in breed prediction.

We report the successful use of an ANN for locus minimization up to nine loci drastically reduces the cost of genotyping of an unknown sample threefold without compromising the prediction accuracy of more than 96% for 22 goat breeds.

Acknowledgements

This work was supported by NAIP, ICAR, Government of India. The kind help of Dr. GPS Raghava, IMTECH, Chandigarh, India, in resolving technical issues of the web server and training is acknowledged. The technical assistance of Jai Bhagwan in maintaining the webserver and AR Paul in designing the server logo is thankfully acknowledged. Authors acknowledge the critical inputs of all the anonymous reviewers and the editor in improvement of the manuscript.

References

Arranz J., Bayon Y. & Primitivo F.S. (2001) Differentiation among Spanish sheep breeds using microsatellites. *Genetics Selection Evolution* **33**, 529–42.

Bjornstad G. & Roed K.H. (2001) Breed demarcation and potential for breed allocation of horse assessed by microsatellite markers. *Animal Genetics* **32**, 59–65.

Bjornstad G. & Roed K.H. (2002) Evaluation of factors affecting individual assignment precision using microsatellite data from horse breeds and simulated breed crosses. *Animal Genetics* **33**, 264–70.

Blott S.C., Williams J.L. & Haley C.S. (1999) Discriminating among cattle breeds using genetic markers. *Heredity* **82**, 613–9.

Dalvit C., De Marchi M. & Cassandro M. (2007) Genetic traceability of livestock products: a review. *Meat Science* **77**, 437–49.

Dixit S.P., Verma N.K., Aggarwal R.A.K., Vyas M.K., Rana J. & Sharma A. (2012) Genetic diversity and relationship among Indian goat breeds based on microsatellite markers. *Small Ruminant Research* **105**, 38–45.

Fan B., Chen Y.Z., Moran C., Zhao S.H., Liu B., Yu M., Zhu M.J., Xiong T.A. & Li K. (2005) Individual-breed assignment analyses in swine populations by using microsatellite marker. *Asian-Australian Journal of Animal Sciences* **11**, 1529–34.

Gotz K. & Thaller G. (1998) Assignment of individuals to populations using microsatellites. *Journal of Animal Breeding and Genetics* **115**, 53–61.

Goudet J. (2002) FSTAT (version 2.9.3.2): a program to estimate and test gene diversities and fixation indices. F-statistics. *Journal of Heredity* **86**, 485–6.

Hamada M., Sato K. & Asai K. (2010) Prediction of RNA secondary structure by maximizing pseudo-expected accuracy. *BMC Bioinformatics* **11**, 586.

Hassoun M.H. (2003). *Fundamentals of Artificial Neural Network*. Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

Hoda A., Hyka G.A., Dunner S. & Obexer-Ruff G. & Econogene Consortium (2011) Genetic diversity of Albanian goat breeds based on microsatellite markers. *Archivos de Zootecnia* **60**, 607–15.

Iquebal M.A., Sarika, Dhanda S.K., Arora V., Dixit S.P., Raghava G.P.S., Rai A. & Kumar D. (2013) Development of a model webserver for breed identification using microsatellite DNA marker. *BMC Genetics* **14**, 118.

Koskinen M.T. (2003) Individual assignment using microsatellite DNA reveals unambiguous breed identification in the domestic dog. *Animal Genetics* **34**, 297–301.

Kumar S., Dixit S.P., Verma N.K., Singh D.K., Pande A., Kumar S., Chander R. & Singh L.B. (2009) Genetic diversity analysis of the Gohilwari breed of Indian goat (*Capra hircus*) using microsatellite markers. *American Journal of Animal and Veterinary Sciences* **4**, 49–57.

MacHugh D.E., Loftus R.T., Cunningham P. & Bradley D.G. (1998) Genetic structure of seven European cattle breeds assessed using 20 microsatellite markers. *Animal Genetics* **29**, 333–40.

Maudet C., Luikart G. & Taberlet P. (2002) Genetic diversity and assignment tests among seven French cattle breeds based on microsatellite DNA analysis. *Journal of Animal Science* **80**, 942–50.

Niu L.L., Li H.B., Ma Y.H. & Du L.X. (2011) Genetic variability and individual assignment of Chinese indigenous sheep popu-

lations (*Ovis aries*) using microsatellites. *Animal Genetics* **43**, 108–11.

Piry S., Alapetite A., Cornuet J.M., Paetkau D., Baudouin L. & Estoup A. (2004) GENECLASS2: a software for genetic assignment and first-generation migrant detection. *Journal of Heredity* **95**, 536–9.

Rai B., Singh M.K. & Singh S.K. (2005) Goats for meat, milk and fibre: a review. *Indian Journal of Animal Sciences* **75**, 349–55.

Sambrook J., Fritsch E.F. & Maniatis T. (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Serrano M., Calvo J.H., Martinez M., Marcos-Carcavilla A., Cuevas J., González C., Jurado J.J. & de Tejada P.D. (2009) Microsatellite based genetic diversity and population structure of the endangered Spanish Guadarrama goat breed. *BMC Genetics* **10**, 61.

Toskinen M.T. & Bredbadka P. (1999) A convenient and efficient microsatellite-based assay for resolving parentage in dogs. *Animal Genetics* **30**, 148–9.

Weir B.S. & Cockerham C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–70.

Yegnanarayama B. (2006). *Artificial Neural Networks*. Prentice Hall of India Private Limited, New Delhi, India.

Yeh F.C., Boyle T., Rongcai Y., Ye Z. & Xian J.M. (1999) POPGENE Version 3.1., http://www.ualberta.ca/~fyeh/index.html.