Supervising Random Forest Using Attribute Interaction Networks

Qinxin Pan[1], Ting Hu[1], James D. Malley[4], Angeline S. Andrew[2,3], Margaret R. Karagas[2,3], and Jason H. Moore[1,2,3]

[1]Department of Genetics, Geisel School of Medicine, Dartmouth College, Hanover, NH 03755, USA
[2]Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Hanover, NH 03755, USA
[3]Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, NH 03755, USA
[4]Division of Computational Bioscience, Center for Information Technology, National Institutes of Health, Bethesda, MD 20892, USA
jason.h.moore@dartmouth.edu

Abstract. Genome-wide association studies (GWAS) have become a powerful and affordable tool to study the genetic variations associated with common human diseases. However, only few of the loci found are associated with a moderate or large increase in disease risk and therefore using GWAS findings to study the underlying biological mechanisms remains a challenge. One possible cause for the "missing heritability" is the gene-gene interactions or epistasis. Several methods have been developed and among them Random Forest (RF) is a popular one. RF has been successfully applied in many studies. However, it is also known to rely on marginal main effects. Meanwhile, networks have become a popular approach for characterizing the space of pairwise interactions systematically, which can be informative for classification problems. In this study, we compared the findings of Mutual Information Network (MIN) to that of RF and observed that the variables identified by the two methods overlap with differences. To integrate advantages of MIN into RF, we proposed a hybrid algorithm, MIN-guided RF (MINGRF), which overlays the neighborhood structure of MIN onto the growth of trees. After comparing MINGRF to the standard RF on a bladder cancer dataset, we conclude that MINGRF produces trees with a better accuracy at a smaller computational cost.

Keywords: Random Forest, Mutual Information Network, Mutual Information Network guided Random Forest, Classification.

1 Introduction

The current strategy for studying the genetic basis of disease susceptibility is to measure millions of single nucleotide polymorphisms (SNPs) across the human genome and test each of them individually for association [14,30]. Genome-wide

L. Vanneschi, W.S. Bush, and M. Giacobini (Eds.): EvoBIO 2013, LNCS 7833, pp. 104–116, 2013.

association studies (GWAS) are based on the idea that genetic variations with alleles common in the population will additively explain much of the heritability of common diseases. As the cost for genome-wide genotyping decreases, the number of GWAS has increased considerably and this approach is now relatively common. The GWAS approach has been successful in that hundreds of new disease-associated SNPs have been reported using rigorous statistical significance and replication criteria [20]. It is anticipated that those SNPs will reveal new pathobiology that will in turn lead to new treatments. While this may be true, few of the loci identified are associated with a moderate or large increase in disease risk and some empirically identified genetic risk factors have been missed [25]. At best, about 20% of the total genetic variance has been explained for a few select common diseases such as the Crohns disease [12]. As a result, many have asked where the missing heritability is [11]. One possibility is that complexities such as gene-gene interactions or epistasis can limit the power of analysis approaches that only consider one SNP at a time [22,23,31].

To faithfully capture the relationships among SNPs several machine learning methods have been considered, including Random Forest (RF) [8,21]. These are, however, engines for making predictions and not necessarily for declaring complexity among the features in the prediction. RF, in particular, is driven by estimating marginal effects in its tree-building process and this is not necessarily a complexity-seeking scheme. RF is one of the most popular ensemble learning methods and has many applications [4,6,10]. A decision tree classifies subjects as case or control by sorting them through a tree from node to node, where each node is a variable with a decision rule that guides that subject through different branches of the tree to a leaf that provides its classification [3]. A RF is a collection of individual decision tree classifiers, where each tree in the forest is trained using a bootstrap sampling of instances (i.e. subjects) from the data, and each variable in the tree is chosen from a random subset of variables. Classification of instances is based upon aggregate voting over all trees in the forest [3]. Although powerful, decision-tree-based methods have one major limitation, that the standard implementations condition on marginal effects [28]. In other words, the algorithm finds the best single variable for the root node before adding additional variables as nodes to the model. This can preclude the detection of epistasis in the absence of significant single SNP effects [27,28,31]. Moreover, multiple variables randomly drawn from the dataset are evaluated and only the best one among them is used for subject sorting. The evaluation of multiple random variables makes RF computationally expensive.

Meanwhile, network science has been used to model interactions and dependencies [1,7,16]. Recently, Hu *et al* [15] proposed Statistical Epistasis Networks (SEN) to characterize the space of pairwise interactions in population-based genetic association studies [18]. In the network, each vertex corresponds to a SNP. An edge linking a pair of vertices corresponds to an interaction between two SNPs. Weights assigned to each SNP and each pair of SNPs quantify how much of the disease status the corresponding SNP and SNPs pair can explain. SEN displays a global representation of all pairwise neighborhood relationship, which

could potentially provide information to supervise classification. However, the network alone does not make a prediction, which makes it hard to interpret.

Given the strengths and weaknesses of RF and networks, we consider a hybrid scheme that has the strengths of both and the weaknesses of neither. Specifically, we embed the unsupervised process of network building into the supervised process of prediction: we use the structure of networks to guide the growing of the forest. In this way, we are able to use the knowledge about variable-variable relationship during the growth of trees, which could potentially make RF less biased towards the marginal main effects, and more efficient by avoiding a random sampling of variables.

The above approach, named Mutual Information Network guided Random Forest (MINGRF), is compared with standard RF on a population-based bladder cancer dataset. The results show that MINGRF produces trees with better accuracies in a shorter runtime.

2 Methods

2.1 Bladder Cancer Dataset

The dataset used in this study consisted of cases of bladder cancer among New Hampshire residents, 25 to 74 years of age, diagnosed from July 1, 1994 to December 31, 2001, and identified in the State Cancer Registry. Controls less than 65 years of age were selected using population lists obtained from the New Hampshire Department of Transportation. Controls 65 years of age and older were chosen from data files provided by the Centers for Medicare & Medicaid Services (CMS) of New Hampshire. This dataset shared a control group with a study of non-melanoma skin cancer in New Hampshire covering an overlapping diagnostic period of July 1, 1993 to June 30, 1995. Additional controls for bladder cancer cases diagnosed from July 1, 1995 to June 30, 1998 were selected with matching age and gender.

Genotyping was performed using the GoldenGate Assay System. The missing value of an individual was filled using the most common genotype of corresponding SNP in the population. The dataset used in our analysis consisted of 491 bladder cancer cases and 71 controls. 1,422 SNPs are included in the dataset. More details on this dataset and the methods are available in [2,17].

2.2 Mutual Information Network

In mathematical terms, a network is a graph, where a graph G consists of a set $V(G)$ of vertices and a set $E(G)$ of edges [24]. In our Mutual Information Networks (MIN), each vertex corresponds to a SNP, and we use v_{A} to denote the vertex corresponding to SNP A. An edge linking a pair of vertices, for instance v_{A} and v_{B}, represents an interaction between SNPs A and B. We first assigned a weight to each pair of SNPs to quantify how much of the disease status the corresponding SNP pair genotypes together explain. In information theoretic terms, the weight corresponds to the two-way *mutual information* [9]. Specifically, the

weight of the edge connecting v_{A} and v_{B} is $I(A, B; C)$, the mutual information of SNPs A and B together with C, the class variable with status *case* or *control*. Intuitively, $I(A, B; C)$ is the reduction in the uncertainty of the class C due to knowledge about SNP A and B's genotypes. Its precise definition is

$$I(A, B; C) = H(C) - H(C|A, B), \tag{1}$$

where $H(C)$ is the *entropy* of C, i.e., the measure of the uncertainty of class C, and $H(C|A, B)$ is the *conditional entropy* of C given knowledge of SNP A and B. Entropy and conditional entropy are defined by

$$H(C) = \sum_{c} p(c) \log \frac{1}{p(c)}, \tag{2}$$

$$H(C|A, B) = \sum_{a,b,c} p(a, b, c) \log \frac{1}{p(c|a, b)}, \tag{3}$$

where $p(c)$ is the probability that an individual has class c, $p(a, b, c)$ is that of having genotype a, b and class c, and $p(c|a, b)$ is that of having class c given the occurrence of genotype a and b together.

Similar to the framework of Statistical Epistasis Network by Hu *et al*, the threshold of including pairwise interactions can be derived systematically by analyzing the topological properties of the networks [15], such as the size of a network, the connectivity of a network (the size of its largest connected component), and its vertex degree distribution. Permutation testing is often used to provide a null distribution of properties of networks built from permuted data. This null distribution can be used to determine the threshold of pairwise strength that mostly distinguishes the real-data network from the permuted-data networks.

2.3 Random Forest

Random Forest is an ensemble learning method. A forest consists of multiple decision tree classifiers and the classification of subjects is based on aggregate voting over all trees.

Specifically, a standard RF procedure takes the following steps [3,6,19]: i) draw *ntree* bootstrap samples from the original data; ii) grow a tree for each bootstrap dataset. At each node of the tree, randomly select *mtry* variables and choose the splitting node that separates cases and controls the best. iii) a tree grows to the largest extent when the number of subjects in a node reaches a minimum *nodesize* and the prediction of that node is decided by the majority class of subjects on that node; iv) aggregate information from the *ntree* trees for new data prediction such as majority voting for classification; v) compute an out-of-bag (oob) accuracy by using the data not in the bootstrap sample [5]. The balanced accuracy, that is the average of sensitivity and specificity, is reported in this study as it is more robust to imbalanced biomedical datasets [29].

2.4 Mutual Information Network Guided Random Forest

We investigate whether MIN can help improve RF by implementing a hybrid algorithm, Mutual Information Network guided Random Forest (MINGRF). To impose the structure of MIN into RF, we implement MINGRF in the following way.

1. When starting building trees, instead of sampling a random set of variables and choosing the one which separates cases and controls best, MINGRF chooses one vertex from the hubs in MIN (vertices that have at least 5 neighbors specified in this study) with a probability proportional to their degrees.
2. While growing the trees, instead of trying a list of variables and choosing the best-case-control-separating one, MINGRF considers all the neighbors of the mother node in MIN which have not been used yet in the building of the current tree and chooses one with probability proportional to the corresponding edge weights. If all neighbors have been used previously, MINGRF chooses one of them with a probability proportional to the edge weights.
3. The growing of a tree continues until the number of samples on a node is smaller than a pre-specified number, i.e. the terminal node size.
4. After the construction of the forest, the quality of trees can be assessed using oob samples in the same manner as standard RF.

In this study, to compare the performances of MINGRF and RF, we use $ntree = 1000$. A wide range of $mtry$ and $nodesize$ are explored in our implementation.

3 Results

3.1 Mutual Information Network

To pick the threshold t at which mutual information network is most different from by chance, we look into the connectivity of the network, i.e. the size of the largest connected component, at decreasing t. Recall that an edge linking SNPs A and B is included in the mutual information network G_t only if their two-way mutual information $I(A, B; C) \geq t$. Accordingly, the networks G_t grow as t decreases.

Figure 1 shows the size of the largest connected component in the network G_t and in the permuted-data networks as t decreases from 0.030 to 0.005 in increments of 0.0001. The largest connected component of G_t grows quickly when t decreases from 0.025 to 0.0156 whereas the largest connected component in the permuted-data networks do not start growing until a smaller value of t is reached. The P value of the largest connected component size, estimated based on permutation testing, is smaller than 0.001 when $t \in [0.0155, 0.0299]$. We choose $G_{0.0156}$ for future study as all 1,422 SNPs were included in the largest connected component for the first time when t reaches 0.0156.

The network $G_{0.0156}$ (Figure 2) has 1,422 vertices and 2,236 edges. As SNP IGF2AS_04 has the strongest main effect, every SNP pair which includes IGF2AS

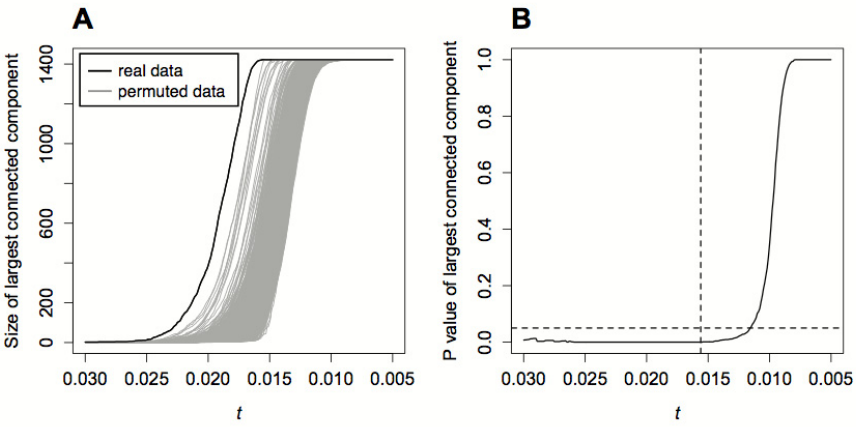

Fig. 1. The size of the largest connected component and its significance in the networks with decreasing threshold t. (A) The size of the largest connected component in G_t and networks of permuted datasets. The black line represents the real-data network G_t and the gray lines represent the networks of 1,000 permuted datasets. The largest connected components include increasingly more vertices as t decreases and eventually include all 1,422 vertices when t reaches 0.0156. (B) The P value of the largest connected component in G_t. P value is estimated as the fraction of networks from permuted datasets whose largest connected component is no smaller than that of G_t. The horizontal dashed line represents P=0.05 and the vertical dashed line represents t=0.0156.

_04 has a relatively high $I(A, B; C)$. Naturally, vertex IGF2AS_04 is connected to every other vertex in $G_{0.0156}$ and has a degree of 1,421. There are 861 vertices which are only connected to IGF2AS_04 (shown on the left) and 560 vertices which have at least one more neighbor besides IGF2AS_04 (shown on the right). Note that there are a large set of edges in which IGF2AS_04 is not involved. In other words, interactions or additive effects that are independent of IGF2AS_04 also contribute to bladder cancer risk, which further indicates the fact that bladder cancer is a complex disease.

3.2 Mutual Information Network and Random Forest Network Comparison

To compare MIN and RF, we ask the research question whether important variables in RF are also identified as important in MIN. Gini importance, which indicates both how often a particular variable is selected for a split and how large its overall discriminative value is for the classification problem under study, is used to quantify the importance of a particular variable in RF. Node degree, defined as the number of neighbors a specific vertex has in the graph, is used to assess the importance of that variable in MIN. Recall that MIN $G_{0.0156}$ is chosen as all 1,422 SNPs are included in the largest connected component when

Fig. 2. Mutual Information Network $G_{0.0156}$. There are 1,422 vertices and 2,236 edges. As vertex IGF2AS_04 (shown in the middle) has strong main effect which contributes to two-way mutual information, it is connected to every other vertex in the graph. For visualization purpose, all edges that connect vertex IGF2AS_04 are shown in gray all the other edges are shown in black. Vertices which are only connected to IGF2AS_04 are shown on the left side and vertices which have at least one more other neighbors besides IGF2AS_04 are shown on the right side. The graph is generated by the software Cytoscape [26].

the threshold $t \leq 0.0156$ (Figure 1). Figure 4A shows a significant correlation between the degree of MIN $G_{0.0156}$ and RF Gini importance (Spearman's rank correlation coefficient ρ=0.495 and $P<2.2 \times 10^{-16}$).

As RF Gini importance is known to be biased towards main effect [27,28], to fully compare MIN and RF with SNP-SNP relationships taken into account, we convert a forest into a RF network and compared it with MIN. Given a forest, we count the occurrence of two SNPs being mother-daughter nodes in all the trees and assign the occurrence as the weight to the edge between the SNP pair. Similar to the threshold-based MIN, an edge is included in the RF network $\hat{G}_t$ only when its weight is no less than a particular threshold t. Figure 3 shows the number of edges and the size of largest connected component in the network of real data and networks of permuted data as t decreases from 4 to 1 in increments of 1. When $t \leq 2$, the networks of permuted data possessed much more edges than that of real data, whereas when $t \geq 3$ the difference became negligible (Figure 3A). The size of largest connected component in networks of permuted data was significantly larger than that of real data when t=2 (Figure 3B). Based on above observations, we chose t=2 as a threshold for later study.

After obtaining a RF network $\hat{G}_2$, we are able to compare its node degree with that of the MIN $G_{0.0156}$. A significant correlation between the degree of MIN $G_{0.0156}$ and that of RF network G_2 is observed with Spearman's rank correlation coefficient ρ=0.483 and $P< 2.2 \times 10^{-16}$ (Figure 4B). Although correlated, the low correlation coefficients indicate the fact that the methods overlap with differences.

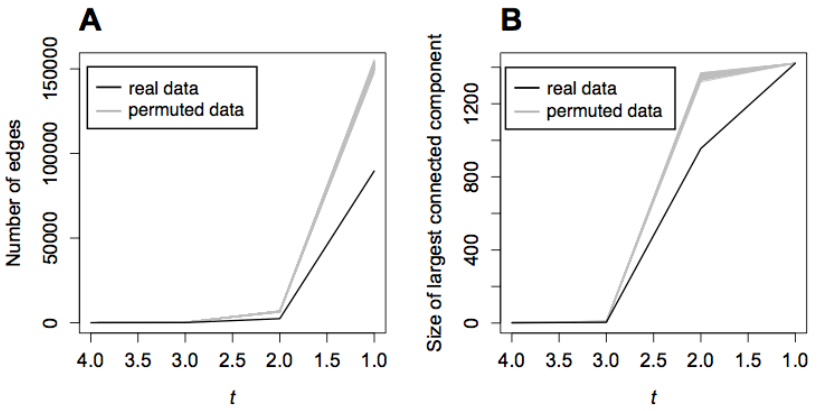

Fig. 3. Random Forest Network growth with decreasing threshold t. (A) Increase in the number of edges. (B) Increase in the size of largest connected component. In both panels, the black line represents $\hat{G}t$ of the real data and the gray lines represent networks of 1,00 permuted datasets. The threshold t, denoted as the times two corresponding SNPs show up in the forest as mother-daughter nodes, decreases from 4 to 1 in increments of 1.

3.3 Mutual Information Network Guided Random Forest

We evaluate the performance of MIN guided RF (MINGRF) using out-of-bag (oob) accuracy and runtime (Figure 5). Recall that RF has three key parameters: the terminal node size, the number of variables randomly sampled at each splitting and the number of trees in the forest. As MINGRF does not randomly sample a set of variables at each splitting, it has only two: the terminal node size and the number of trees. We thoroughly compare their performances under a wide range of different parameters and find that MINGRF always has better oob accuracy than RF (Figure 5A and 5B). Although RF oob accuracy increases as the number of variables sampled increases, the runtime of RF also increases accordingly (Figure 5B and 5C). The runtime of RF is shorter than that of MINGRF when the number of variables sampled is small, but it exceeds that of MINGRF quickly when the number of variables sampled starts to increase (Figure 5C).

4 Discussion

In this article, we compare the findings of Mutual Information Network and Random Forest in a bladder cancer dataset and observe both similarities and differences. The differences allow the potential for improvement which might be achieved by combining the two methods. Encouraged by these findings, to further integrate the advantages of MIN into RF, we propose a hybrid algorithm by imposing the neighborhood relationship of MIN into the tree construction of RF, i.e. MIN guided RF (MINGRF). Usually, RF randomly samples a list of variables

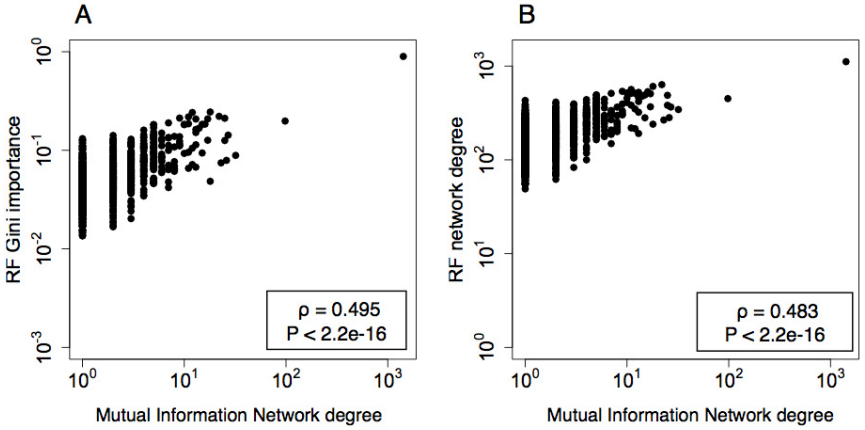

Fig. 4. Comparison of Mutual Information Network $G_{0.0156}$ and Random Forest. (A) Correlation of Mutual Information Network $G_{0.0156}$ node degree and RF Gini importance. (B) Correlation of Mutual Information Network $G_{0.0156}$ node degree and RF Network G_2 node degree. In both panels, Spearman's rank correlation coefficient ρ and the corresponding P value are reported.

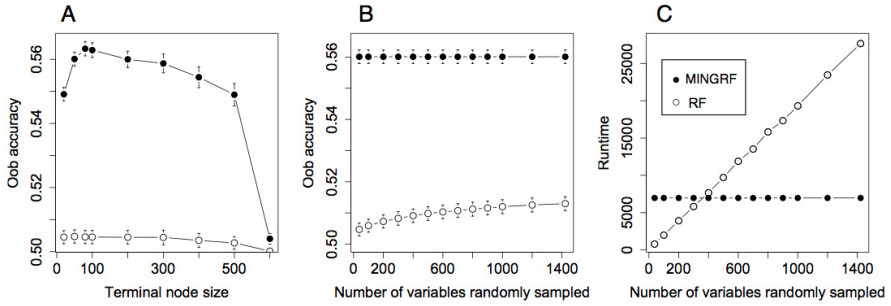

Fig. 5. Performance of Mutual Information Network Guided RF and RF. (A) Balanced oob accuracy shown as a function of terminal node size. (B) Balanced oob accuracy shown as a function of number of variables randomly sampled at each splitting. (C) Run time shown as a function of number of variables randomly sampled at each splitting. Data represent the mean of 1,000 independent replications and error-bars denote 95% confidence intervals.

and greedily chooses the one which separates cases and controls the best at each splitting. This process relies on the marginal main effect and is computationally expensive. In contrast, MINGRF takes advantage of the pairwise interaction landscape and put the strong pairs in MIN adjacent to each other in the tree, which not only takes SNP-SNP relationship (i.e. interaction, additive effect) into account but also improves computational efficiency. We find that the trees

produced by MINGRF have better oob accuracies and shorter runtime for a broad range of different parameters.

The comparison of MIN and RF leads to a few interesting observations. First, the MIN of real data has more vertices on the largest connected component than permuted data (Figure 1). As the edge number of real data and that of permuted data are not significantly different (data not shown), the difference of the largest connected component size is most likely caused by the clustering of interacting SNPs, in other words, the existence of hubs. Second, the RF networks of permuted datasets possess more edges and more vertices on the largest connected components (Figure 3). This could be partially explained by the fact that in a permuted dataset without real biological signals, it is hard to separate cases and controls and consequently RF learns random noises. Thus, it takes more steps of splitting to reach a certain terminal node size in a permuted dataset than in the real dataset, which leads to more edges in the permuted-data networks. The clear discrepancy between RF network of real data and that of permuted data indicates that our approach of constructing RF network captures the characteristic of the dataset and thus the RF network is comparable with MIN. Third, the positive correlation observed between MIN and RF indicate that the two methods identify similar variables. However, given the low correlation coefficients, their difference is definitely not negligible (Figure 4). There are a few possible reasons for the difference: i) MIN is deterministic whereas the sampling process in RF can introduce stochastic noises. Therefore, for variables with low MIN degree, their RF Gini importances can vary a lot; ii) MIN considers SNP-SNP relationships while RF mostly relies on marginal main-effect. SNP pairs which are in interaction without strong marginal main effects would be captured by MIN but not necessarily by RF.

Based on the above observations, we implement a hybrid algorithm called Mutual Information Network guided Random Forest (MINGRF). The goal of designing this algorithm is to refine the random sampling process of RF and consequently improve both the tree quality and runtime. MINGRF has many advantages. i) The choice of significant MIN does not rely on the significance of each pair of SNPs, instead it describes the point when the network as a whole system is most significant; ii) MINGRF takes advantage of the relationship between two variables and is less biased towards main effects; iii) MINGRF produces trees with better accuracies, which could be informative to propose a biological hypothesis; iv) MINGRF avoids the random sampling of variables at each splitting, which makes the construction of trees more efficient.

Among the limitations of this approach, an important one is that as MIN only captures pairwise SNP-SNP relationships, and higher order interactions might be overlooked. Moreover, with the growth of trees in the forest, RF identifies variables which separate cases and controls in a small subset of samples falling on the corresponding node, which encourages the detection of heterogeneity. But MINGRF finds variables essential for the whole population due to the way we construct MIN, which makes it not very useful for heterogeneity. As an alternative we could consider finding interacting features within the subpopulations,

to build networks that are possibly group dependent. We could then test for differences between the two sets of detected networks. There are simple matrix methods for this. If the group-based networks are declared similar it makes sense to declare them as valid population networks. Otherwise group differences would be captured by the separate networks, and these would be adapted to heterogeneity.

Future work includes comparing MINGRF with RF more thoroughly using cross fold validation, tree consistency etc. As the construction process is more transparent in MINGRF, we expect to get more interpretable models. Moreover, we are also interested in studying the top variables identified by MINGRF and RF. Whether the top variables are truly in interactions can be tested using explicit test [13]. Whether the usage of local neighborhood relationships in MIN will help the findings of higher order interactions will also be interesting to investigate. As MINGRF uses information about SNP-SNP relationships, we expect MINGRF to detect interactions which are usually overlooked by standard RF [28].

In conclusion, we compare two methods, Mutual Information Network (MIN) and Random Forest (RF), and observed both similarities and differences. MIN captures the two-way interaction landscape well yet can not give a prediction itself. On the other hand, RF is powerful in classification problems but also is known to be biased towards marginal main effects. After a thorough comparison, we propose a novel algorithm MIN guided RF (MINRF) and test it on a bladder cancer dataset. We conclude that MINRF yields decision trees with better accuracies at a lower computational cost.

Acknowledgments. This work was supported by NIH grants R01-LM009012, R01-LM010098, and R01-AI59694.

References

1. Andrei, A., Kendziorski, C.: An efficient method for identifying statistical interactors in gene association networks. Biostatistics 10(4), 706–718 (2009)
2. Andrew, A.S., Nelson, H.H., Kelsey, K.T., Moore, J.H., Meng, A.C., Casella, D.P., Tosteson, T.D., Schned, A.R., Karagas, M.R.: Concordance of multiple analytical approaches demonstrates a complex relationship between dna repair gene snps, smoking and bladder cancer susceptibility. Carcinogenesis 27(5), 1030–1037 (2006)
3. Breiman, L.: Random forests. Mach. Learn. 45, 5–32 (2001)
4. Bureau, A., Dupuis, J., Falls, K., Lunetta, K.L., Hayward, B., Keith, T.P., Van Eerdewegh, P.: Identifying snps predictive of phenotype using random forests. Genet. Epidemiol. 28(2), 171–182 (2005)
5. Bylander, T.: Estimating generalization error on two-class datasets using out-of-bag estimates. Machine Learning 48, 287–297 (2002)
6. Chen, X., Ishwaran, H.: Random forests for genomic data analysis. Genomics 99(6), 323–329 (2012)
7. Chu, J.H., Weiss, S.T., Carey, V.J., Raby, B.A.: A graphical model approach for inferring large-scale networks integrating gene expression and genetic polymorphism. BMC Syst. Biol. 3, 55 (2009)

8. Cordell, H.J.: Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum. Mol. Genet. 11(20), 2463–2468 (2002)
9. Cover, T.M., Thomas, J.A.: Elements of information theory, 2nd edn. Wiley (2006)
10. Diaz-Uriarte, R., Alvarez de Andres, S.: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3 (2006)
11. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., Nadeau, J.H.: Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. 11(6), 446–450 (2010)
12. Franke, A., McGovern, D.P.B., Barrett, J.C., Wang, K., Radford-Smith, G.: Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. Nat. Genet. 42(12), 1118–1125 (2010)
13. Greene, C.S., Himmelstein, D.S., Nelson, H.H., Kelsey, K.T., Williams, S.M., Andrew, A.S., Karagas, M.R., Moore, J.H.: Enabling personal genomics with an explicit test of epistasis. In: Pac. Symp. Biocomput., pp. 327–336 (2010)
14. Hirschhorn, J.N., Daly, M.J.: Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. 6(2), 95–108 (2005)
15. Hu, T., Sinnott-Armstrong, N.A., Kiralis, J.W., Andrew, A.S., Karagas, M.R., Moore, J.H.: Characterizing genetic interactions in human disease association studies using statistical epistasis networks. BMC Bioinformatics 12(364) (2011)
16. Ideker, T., Sharan, R.: Protein networks in disease. Genome Res. 18(4), 644–652 (2008)
17. Karagas, M.R., Tosteson, T.D., Blum, J., Morris, J.S., Baron, J.A., Klaue, B.: Design of an epidemiologic study of drinking water arsenic exposure and skin and bladdder cancer risk in a U.S. population. Environ. Health Perspect. 106(suppl. 4), 1047–1050 (1998)
18. Lavender, N.A., Rogers, E.N., Yeyeodu, S., Rudd, J., Hu, T., Zhang, J., Brock, G.N., Kimbro, K.S., Moore, J.H., Hein, D.W., Kidd, L.C.R.: Interaction among apoptosis-associated sequence variants and joint effects on aggressive prostate cancer. BMC Med. Genomics 5, 11 (2012)
19. Malley, J.D., Kruppa, J., Dasgupta, A., Malley, K.G., Ziegler, A.: Probability machines: Consistent probability estimation using nonparametric learning machines. Methods Inf. Med. 10(51), 74–81 (2011)
20. Manolio, T.A.: Genomewide association studies and assessment of risk of disease. New England Journal of Medicine 363(2), 166–176 (2010)
21. McKinney, B.A., Reif, D.M., Ritchie, M.D., Moore, J.H.: Machine learning for detecting gene-gene interactions: a review. Appl. Bioinformatics 5(2), 77–88 (2006)
22. Moore, J.H., Asselbergs, F.W., Williams, S.M.: Bioinformatics challenges for genome-wide association studies. Bioinformatics 26(4), 445–455 (2010)
23. Moore, J.H., Williams, S.M.: Epistasis and its implications for personal genetics. Am. J. Hum. Genet. 85(3), 309–320 (2009)
24. Newman, M.: Networks: An introduction. Oxford University Press (2010)
25. Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F., Moore, J.H.: Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am. J. Hum. Genet. 69(1), 138–147 (2001)
26. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13(11), 2498–2504 (2003)

27. Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A.: Conditional variable importance for random forests. BMC Bioinformatics 9(307) (2008), doi:10.1186/1471–2105–9–307
28. Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T.: Bia in random forest variable importance measures: Illustration, sources and a solution. BMC Bioinformatics 8(25) (2007), doi:10.1186/1471–2105–8–25
29. Velez, D.R., White, B.C., Motsinger, A.A., Bush, W.S., Ritchie, M.D., Williams, S.M., Moore, J.H.: A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genetic Epidemiology 31(4), 306–315 (2007)
30. Wang, W.Y.S., Barratt, B.J., Clayton, D.G., Todd, J.A.: Genome-wide association studies: theoretical and practical concerns. Nat. Rev. Genet. 6(2), 109–118 (2005)
31. Williams, S.M., Canter, J.A., Crawford, D.C., Moore, J.H., Ritchie, M.D., Haines, J.L.: Problems with genome-wide association studies. Science 316(5833), 1840–1842 (2007)