ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/221136502

Feature Selection using a Random Forests Classifier for the Integrated Analysis of Multiple Data Types

Conference Paper · September 2006

DOI: 10.1109/CIBCB.2006.330987 · Source: DBLP

CITATIONS
57

READS
2,936

5 authors, including:

David M Reif
North Carolina State University
185 PUBLICATIONS **6,565** CITATIONS

SEE PROFILE

Brett Mckinney
University of Tulsa
102 PUBLICATIONS **1,883** CITATIONS

SEE PROFILE

James E Crowe
Vanderbilt University Medical Center, Nashville, TN, United States
778 PUBLICATIONS **20,505** CITATIONS

SEE PROFILE

Jason H Moore
University of Pennsylvania
885 PUBLICATIONS **31,240** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Antibody modeling, Antibody design and Antigen-Antibody interactions View project

Hybrid biclustering algorithms View project

All content following this page was uploaded by James E Crowe on 27 May 2014.

The user has requested enhancement of the downloaded file.

Feature Selection using a Random Forests Classifier for the Integrated Analysis of Multiple Data Types

David M. Reif[1,2], Alison A. Motsinger[1], Brett A. McKinney[1,2,3], James E. Crowe Jr.[3], Jason H. Moore[2]
{reif, motsinger, mckinney}@chgr.mc.vanderbilt.edu, james.e.crowe@vanderbilt.edu, jason.h.moore@dartmouth.edu

[1]Center for Human Genetics Research, Dept. of Molecular Physiology & Biophysics, Vanderbilt University, Nashville, TN, USA

[2]Computational Genetics Laboratory, Dept. of Genetics, Dartmouth Medical School, Lebanon, NH, USA

[3]Program in Vaccine Sciences, Dept. of Pediatrics, Vanderbilt University, Nashville, TN, USA

ABSTRACT

Complex clinical phenotypes arise from the concerted interactions among the myriad components of a biological system. Therefore, comprehensive models can only be developed through the integrated study of multiple types of experimental data gathered from the system in question. The Random Forests™ (RF) method is adept at identifying relevant features having only slight main effects in high-dimensional data. This method is well-suited to integrated analysis, as relevant attributes may be selected from categorical or continuous data, and there may be interactions across data types. RF is a natural approach for studying gene-gene, gene-protein, or protein-protein interactions because importance scores for particular attributes take interactions into account. Thus, Random Forests is a promising solution to the analysis challenge posed by high-dimensional datasets including interactions among attributes of different types. In this study, we characterize the performance of RF on a range of simulated genetic and/or proteomic datasets. We compare the performance of RF in identifying relevant attributes when given genetic data alone, proteomic data alone, or a combined dataset of genetic plus proteomic data. Our results indicate that utilizing multiple data types is beneficial when the disease model is complex and the phenotypic outcome-associated data type is unknown. The results of this study also show that RF is adept at identifying relevant features in high-dimensional data with small main effects and low heritability.

Keywords

Random Forests™, gene-gene interactions, feature selection, multiple data types, data integration.

1. INTRODUCTION

Adverse drug reaction is one of the leading causes of hospitalizations in the Unites States. For example, in 1994 alone, adverse drug reactions accounted for more than 2.2 million serious hospitalizations [1]. Currently, there is no definitive way to determine how a person will respond to a medication—limiting pharmaceutical development to a "one size fits all" system. This system allows for the development of drugs to which the "typical" patient will respond, but one size does not necessarily fit all, sometimes with dire consequences. The need to screen patients for biomarkers predictive of response *a priori* to prevent adverse reactions has created a subspecialty within the field of human genetics known as pharmacogenomics.

The goal of pharmacogenomics is the identification and characterization of genes that predict drug response [2]. Due to the inherent complexity of the response phenotype, it is hypothesized that patient outcome is largely dependent upon interactions among genes and the environment. These nonlinear genetic interactions, known as epistasis, quickly diminish the applicability of traditional statistical methods. Taken together with the current explosion of genetic information as the field pushes towards genome-wide association studies, epistasis presents analytical challenges of an enormous combinatorial magnitude [3;4]. Traditional parametric analysis methods can be overwhelmed by datasets having huge numbers of attributes yet few samples. In response to the complex nature of current genetic studies, a number of novel statistical and computational methods have been developed [5-9].

Even with suitable analytical methodology, considering experimental information gathered from only one type of biological data will not permit the capture of the enormous complexity of systemic response phenotypes. Systems biology seeks to integrate multiple levels of information to understand how biological systems function [10]. By studying the relationships and interactions between various parts of a biological system,

1-4244-0623-4/06/$20.00 ©2006 IEEE.

a more comprehensive model can be developed. Furthermore, because biology operates through a hierarchy of levels, incorporating data from multiple levels can provide surrogate data to fill gaps from any one biological level, and the partial redundancy between levels can further mitigate methodological unreliability [11].

For pharmacogenomic studies, an initial systems biology approach might measure variation in both genes and proteins in a patient to identify biomarkers that predict response to a given drug. While there is intuitive appeal to such a strategy, adding pieces of information on different scales of measurement (*i.e.* continuous proteomic data as well as categorical genetic data) creates additional analytical challenges. Therefore, appropriate computational analysis methods must not only traverse large numbers of input variables, but will also need to handle diverse data types.

One such computational method is the Random Forests (RF) approach [12]. RF is a machine learning technique that builds a forest of classification trees wherein each tree is grown on a bootstrap sample of the data, and the attribute at each tree node is selected from a random subset of all attributes. The final classification of an individual is determined by voting over all trees in the forest. There are many advantages of the RF method that make it an ideal approach for the analysis of diverse biological data in pharmacogenomic studies. First, it can handle a large number of input attributes—both qualitative (*e.g.* Single Nucleotide Polymorphisms, or "SNPs") and quantitative (*e.g.* microarray expression levels or data from high-throughput proteomic technologies). Second, it estimates the relative importance of attributes in determining classification, thus providing a metric for feature selection. Third, RF produces a highly accurate classifier with an internal unbiased estimate of generalizability during the forest building process. Fourth, RF is fairly robust in the presence of etiological heterogeneity and relatively high amounts of missing data [13]. Finally, and of increasing importance as the number of input variables increases, learning is fast and computation time is modest even for very large datasets [14].

In the current study, we use simulated data to investigate the potential of using a RF approach for the combined analysis of both genetic and proteomic data gathered in a study of adverse events associated with trials of a new smallpox vaccine [15;16]. The simulations are based on data collected from recent clinical trials of the Aventis-Pasteur Smallpox Vaccine (APSV), in which a significant proportion of vaccinees suffered systemic adverse events (AEs)—including fever, lymphadenopathy, and generalized rash. The data include genotypes at 1442 SNPs and measured circulating levels of 108 immunological proteins. This dataset was chosen for its complex phenotype, the large number of attributes, and the multiple types of data collected. By using the actual data collected as the basis for our simulations, we reduce the number of over-simplifying assumptions and hope to better model the complexity inherent in real data. Because adverse reaction to vaccination is a complex phenotype, it is likely due to the coordinated action of multiple biological factors. Therefore, our simulated outcome (adverse event) models involve attribute interactions with only slight main effects.

In this study, we evaluate the ability of RF to detect outcome-associated simulated attributes by analyzing genetic data alone, proteomic data alone, or combined genetic and proteomic data. We address several questions with this study. First, in order to address the unresolved issue of where to set the importance cutoff for relevant attributes [13], can we set an appropriate threshold for the calculated RF importance relative to all attributes in the particular dataset analyzed that includes our simulated functional attributes? Second, how does RF perform when given different types of simulated biological data as input? Third, is there a relationship between the degree of informational redundancy and the ability of RF to select proteomic attributes related to the functional genetic attributes? Fourth, are there situations in which the analysis of multiple data types proves beneficial? In brief, our results indicate that utilizing multiple data types is beneficial when the disease model is complex and the outcome-associated data type is unknown. Importantly, using RF, we do not observe any significant *dis*advantage to an analysis strategy integrating both data types.

2. METHODS

2.1 Random Forests

The Random Forests method uses a collection of decision tree classifiers, where each tree in the forest has been trained using a bootstrap sample of individuals from the data, and each split attribute in the tree is chosen from among a random subset of attributes. Classification of individuals is based upon aggregate voting over all trees in the forest.

Each tree in the forest is constructed as follows from data having N individuals and M explanatory attributes:

1. Choose a training sample by selecting N individuals, with replacement, from the entire dataset.
2. At each node in the tree, randomly select m attributes from the entire set of M attributes in the data. The absolute magnitude of m is a function of the number of attributes in the dataset and remains constant throughout the forest building process.
3. Choose the best split at the current node from among the subset of m attributes selected above.
4. Iterate the second and third steps until the tree is fully grown (no pruning).

Repetition of this algorithm yields a forest of trees, each of which have been trained on bootstrap samples of individuals (see Fig. 1). Thus, for a given tree, certain individuals will have been left out during training.

Prediction error and attribute importance is estimated from these "out-of-bag" individuals.

The out-of-bag (unseen) individuals are used to estimate the importance of particular attributes according to the following logic: If randomly permuting values of a particular attribute does *not* affect the predictive ability of trees on out-of-bag samples, that attribute is assigned a low importance score. If, however, randomly permuting the values of a particular attribute drastically impairs the ability of trees to correctly predict the class of out-of-bag samples, then the importance score of that attribute will be high. By running out-of-bag samples down entire trees during the permutation procedure, attribute interactions are taken into account when calculating importance scores, since class is assigned in the context of other attribute nodes in the tree.

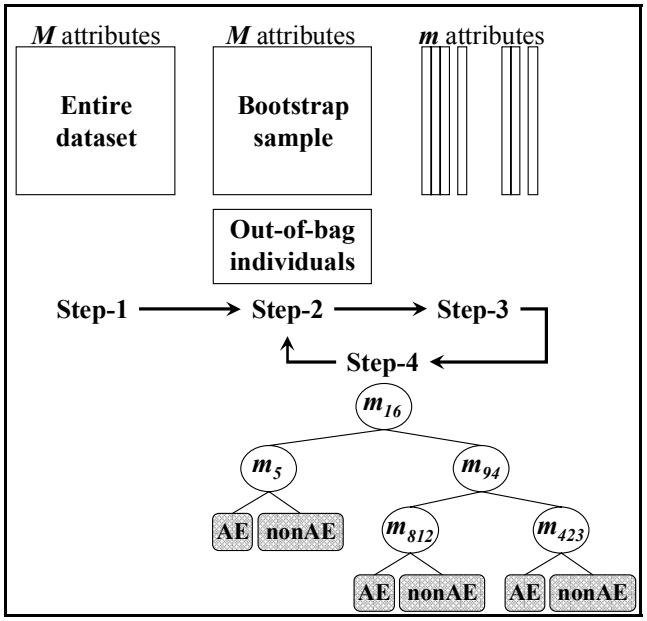

Figure 1: Construction of individual trees using the Random Forest method from a full dataset of *N* individuals and *M* attributes. The steps correspond to those described in Section 2.1.

The recursive partitioning trees comprising a RF provide an explicit representation of attribute interaction that is readily applicable to the study of interactions among multiple data types [17;18]. These models may uncover interactions among genes, proteins, and/or environmental factors that do not exhibit strong marginal effects. Additionally, tree methods are suited to dealing with certain types of genetic heterogeneity, since splits near the root node define separate model subsets in the data. Random Forests capitalize on the solid benefits of decision trees and have demonstrated excellent predictive performance when the forest is diverse (*i.e.* trees are not highly correlated with each other) and composed of individually strong classifier trees [12;19]. The RF method is a natural approach for studying gene-gene, gene-protein, or protein-protein interactions because importance scores for particular attributes take interactions into account without demanding a pre-specified model [20].

2.2 Data Simulation

Simulation studies were designed to assess whether a Random Forests classifier is able to select the appropriate (outcome-associated) attributes from datasets consisting of categorical genetic (SNP) attributes, continuous proteomic (cytokine) attributes, or both. The results of this study will be used to develop an analysis strategy that effectively combines information gathered on diverse biological data types for the vaccine trial described below.

As mentioned previously, the simulations are based on data collected from recent clinical trials of the Aventis-Pasteur Smallpox Vaccine (APSV), where a high proportion of vaccinees suffered systemic adverse events (AEs). These AEs included fever, lymphadenopathy, and generalized rash. The data collected include genotypes at 1442 SNPs (selected from genomic regions within or near candidate genes) and circulating levels of 108 immunological proteins (cytokines). For the APSV data, some proteomic attributes are also represented by genetic data in the corresponding gene. Thus, there is biological overlap between the two data types. Following the protocol described below, the simulated datasets mirrored the actual (APSV) trial data in terms of allele frequencies, SNP distribution across proteins, case (AE)/control (non-AE) ratio, potential patterns of linkage disequilibrium between SNPs, covariance structure across protein levels, etc.

To create simulated data reflecting the complex properties of that collected for the APSV study, those data were used as the basis for the simulations. First, the AE status was stripped from the APSV data. Next, a new AE status was assigned according to genetic attributes in the data consistent with our simulated genetic models and maintaining the overall case/control (AE/nonAE) ratio. Then, to represent the biological transfer of information between genes and proteins, proteomic attributes related to the functional genetic attributes were added. The related proteomic attributes simulate a range of gene→protein information transfer proportions. For example, to simulate a functional (outcome-associated) genetic attribute that is represented by the corresponding protein in the proteomic data, a related proteomic attribute is added to the proteomic data. However, to account for biological variation between genotype and protein level, the functional genetic attribute is only responsible for a portion of the variation in protein level for the related attribute (see Fig. 2). Thus, information is not transferred between related attributes with perfect fidelity.

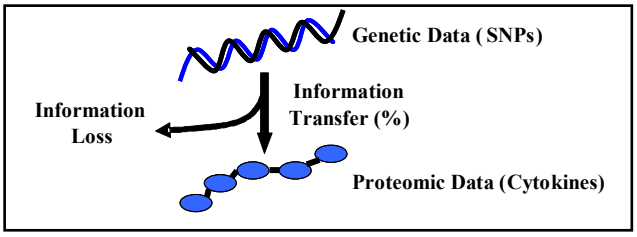

Figure 2: Information transfer between simulated genetic and proteomic attributes. For a particular attribute, the information transfer is the proportion (%) of variation in the simulated proteomic attribute explained by the corresponding genetic attribute.

Penetrance functions are used to represent our partially epistatic genetic models. As in Table 1, penetrance defines the probability of experiencing an adverse event given a particular genotype combination. For these models, two genetic attributes ($Genetic_A$ + $Genetic_B$) have a joint (epistatic) effect upon outcome class, and each attribute also has a very slight marginal effect (*M*) above the population prevalence (*K*). For a particular combination of genotypes at *i*=$Genetic_A$ and *j*=$Genetic_B$, the probability of belonging to the outcome class AE = f_{ij} in Table 1. A range of heritability values was selected for our simulations, including 10%, 20%, and 40%. Roughly, heritability is the proportion of the total variation in outcome that is due to genetic effects. Although the heritability values used here translate to weak signals in the data, these values would classify as low to moderate genetic effects. Since there is scant data relating adverse events after vaccination with APSV to serum proteomic data or SNP data, heritability values in the low- to mid-range of those estimated for common complex phenotypes were used in these simulations. For a more thorough explanation of the heritability calculations used in this study, see [21]. An example of the penetrance functions used for the models generated in this study is given in Table 2 (others available from the authors upon request).

Datasets with a range of genetic→proteomic information transfer (see Fig. 2) were created for each genetic model. For each combination defined in Table 3 by a genetic model heritability (10%, 20%, 40%), a proportion of genetic→proteomic information transfer (15%, 30%, 45%, 60%, 75%, 90%), and a data type (Genetic, Proteomic, Genetic+Proteomic), 100 datasets were simulated for analysis, resulting in 5400 total datasets.

Table 1. Penetrance function for a model of AE status associated with two functional genetic attributes: *A* and *B*.

Genetic Attribute A	Genetic Attribute B: BB	Bb	Bb	
AA	f_{11}	f_{12}	f_{13}	M_{A1}
Aa	f_{21}	f_{22}	f_{23}	M_{A2}
aa	f_{31}	f_{32}	f_{33}	M_{A3}
	M_{B1}	M_{B2}	M_{B3}	***K***

Table 2. Example penetrance function for a simulated genetic AE model with 10% heritability. Allele frequencies for each attribute are equal ($p = q = 0.5$).

Genetic Attribute A	Genetic Attribute B: BB	Bb	bb	
AA	0	0	0	M_{A1}
Aa	0	0.2	0.2	M_{A2}
aa	0	0.2	0.2	M_{A3}
	M_{B1}	M_{B2}	M_{B3}	***K***

Table 3: Overview of simulated datasets. For each combination of genetic heritability and genetic→proteomic information transfer, 100 datasets were simulated, each containing one of the following data types: Genetic data alone = G; Proteomic data alone = P; Genetic + Proteomic data combined = GP.

Genetic Heritability	Genetic-Proteomic Information Transfer: 15%	30%	45%	60%	75%	90%
10%	G,P,GP	G,P,GP	G,P,GP	G,P,GP	G,P,GP	G,P,GP
20%	G,P,GP	G,P,GP	G,P,GP	G,P,GP	G,P,GP	G,P,GP
40%	G,P,GP	G,P,GP	G,P,GP	G,P,GP	G,P,GP	G,P,GP

2.3 Data Analysis

Analysis was performed using the freely available R package randomForest [22;23]. This package is based on the original Fortran code available at [24].

Given a dataset having a particular combination of genetic model heritability and genetic→proteomic information transfer (see Table 3), RF was used to analyze datasets containing each simulated biological data type separately and in parallel. Genetic attributes were treated as categorical while proteomic attributes were treated as continuous values. For each of the 100 genetic, proteomic, or combined datasets, forests comprised of 10,000 trees were grown. Attribute importance was calculated using the out-of-bag permutation test. The

relative importance (rank) of functional genetic attributes and related proteomic attributes was determined from the mean decrease in Gini index using the out-of-bag permutation testing procedure (The relative importance determined from the mean decrease in classification accuracy produced statistically similar results) [17].

3. RESULTS

Fig. 3 shows the relative importance rank (expressed as a percentile) of the two functional genetic attributes calculated by the RF over all datasets. Each data point on the graph represents the mean relative importance rank calculated over 100 datasets, with the bars representing 95% confidence intervals about the mean. This figure demonstrates several important trends regarding the relative importance of the functional genetic variables with the three possible combinations of data types analyzed (Genetic alone = G, Proteomic alone = P, Genetic + Proteomic combined = GP). Analyzing the genetic data alone consistently demonstrated the highest relative importance for the functional genetic attributes. Analyzing the combined genetic + proteomic data demonstrated relative importance that was very near to that of the genetic alone. This slight discrepancy may be due to the increased number of noise attributes (the combined dataset has 1550 attributes while the genetic data alone has only 1442). It is interesting to note that as the heritability of the model increases, the gap in functional attribute importance between the genetic and combined analyses narrows. Of course, regardless of heritability or genetic→proteomic information transfer, analyzing the proteomic data alone makes it impossible to identify the correct genetic attributes since they are not present in the proteomic datasets. Also, as expected, the relative importance of the genetic attributes is not influenced by the amount of information transfer between genetic and proteomic data.

Additionally, it is clear from Fig. 3 that as the heritability of the model increases (across the panels from 10-20-40%), the relative rank of the functional genetic attributes increases. This is expected, since increased heritability increases the signal strength in the data. It is also important to note that even at the lowest heritability simulated (10%), RF successfully identifies the functional variables as relatively important (above the 80th percentile for all models).

Fig. 4 shows the RF relative importance rank of the proteomic variables related to the functional genetic variables (by the % information transfer given along the horizontal axis). Again, each data point represents the mean relative importance rank of the related proteomic attributes calculated over 100 datasets, with the bars representing 95% confidence intervals about the mean. The results are shown for all models, and several significant trends are clear. As expected, when just the genetic datasets are analyzed, it is impossible to identify any proteomic variables as important since they are excluded from those data. Also apparent from Fig. 4 are the wider confidence intervals associated with analysis of the proteomic datasets alone.

As in Fig. 3, increased heritability of the underlying genetic models generally increases the relative importance of outcome-associated attributes (which are the related proteomic attributes in Fig. 4). Unlike the relative importance of genetic attributes considered in Fig. 3, where the results were unaffected by the amount of information transfer between the genomic and proteomic

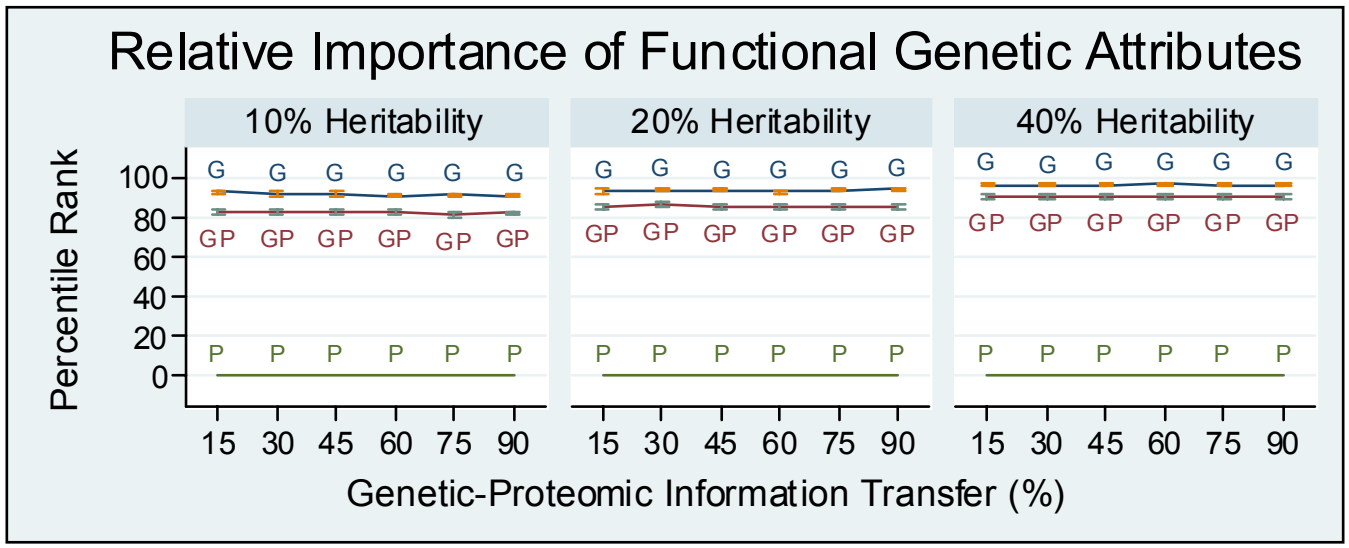

Figure 3: Relative importance of functional genetic outcome-associated attributes for each data type analyzed: Genetic data alone = G; Proteomic data alone = P; Genetic data + Proteomic data combined = GP. Each labeled point represents the mean (plus 95% confidence interval) importance over 100 datasets. Note: the functional genetic attributes are not present in datasets comprised of only proteomic data (P).

data, when considering the related proteomic attributes in Fig. 4, it is clear that the degree of relatedness between the functional genetic attributes and the related proteomic attributes (information transfer) exerts significant influence over the relative importance. This trend is very pronounced in the analysis of the proteomic data alone. As the information transfer increases, the relative importance of the related proteomic attributes increases. The same is true, although to a lesser degree, for the combined genetic + proteomic analyses. Since the disease models are genetic, it is intuitive that as the amount of information transfer between genetic and proteomic attributes increases, the stronger the signal in the proteomic data.

The most striking trend shown in Figure 4 is the large difference between the proteomic and the combined genetic + proteomic analysis strategies. The combined genetic + proteomic analysis strategy is substantially more successful at identifying the related proteomic attributes as important than analysis of the proteomic data alone, especially for models with lower heritability and information transfer. This performance gap may arise out of the partially epistatic nature of the models and the stochastic nature of the RF methodology. Considering models with only slight marginal effects, in order for RF to assign high attribute importance scores, trees must consistently contain both of the relevant interacting attributes. For the combined dataset (containing two functional genetic attributes and two related proteomic attributes), there are more opportunities to choose one of the interacting relevant attributes nearer the root of the tree and then choose the complementary attribute at subsequent splits than for the proteomic data alone (containing only two related proteomic attributes). The performance gap between genetic versus combined datasets in identifying relevant proteomic attributes narrows as both information transfer and heritability increase.

4. DISCUSSION

The results of this study demonstrate that there is a marked advantage to an integrated analysis approach incorporating multiple data types. While the genetic analysis was appropriate for identifying the functional genetic features, the combined strategy analyzing both genetic and proteomic data performed nearly as well at identifying functional genetic attributes and provides another distinct advantage—the identification of important related proteomic variables. This property would be beneficial in situations where the functional outcome-associated data type is unknown or not appropriately measured. For example, our simulated models are not determined by protein *abundance*, as is often measured experimentally. Instead, our simulations represent a situation wherein genotype codes for some unmeasured proteomic aspect (*e.g.* enzymatic activity) that determines phenotype. Still, if protein abundance is also related to genotype, even with some loss of information, the proteomic data can be analytically useful. The convergence of genetic and related proteomic attributes receiving high importance scores could serve as a strategy for limiting false positive results. Also, including multiple data types has the intangible advantage of allowing for better biological interpretation of a resulting model. These results show no substantial *dis*advantage to the joint analysis of multiple data types.

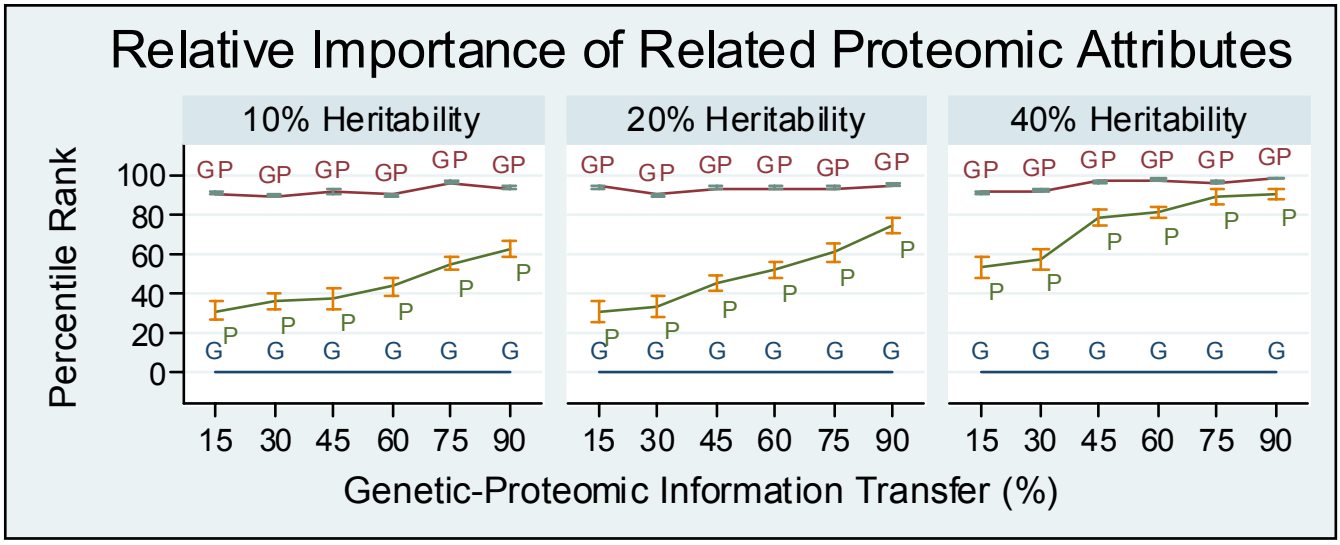

Figure 4: Relative importance of proteomic attributes related (according to the amount of genetic-proteomic information transfer along the horizontal axis) to functional genetic attributes for each data type analyzed: Genetic alone = G, Proteomic alone = P, Genetic + Proteomic combined = GP. Each labeled point represents the mean (plus 95% confidence interval) importance over 100 datasets. Note: functional proteomic attributes are not present in datasets comprised of only genetic data (G).

With respect to setting an appropriate cutoff for selection of relevant features using RF, our results indicate that the choice of threshold depends upon the strength of the signal in the data. It appears that the importance threshold may need to be relaxed to identify relevant attributes in datasets with low signal and a low degree of information transfer between related data types. However, RF seems largely robust to the addition of noise variables—so long as relevant attributes are present in the data. The results of this study also show that RF is adept at identifying relevant features in high-dimensional data containing attributes on multiple scales of measurement. RF identifies features with small marginal effects and low heritability. Relevant attributes may be selected from either data type, and there may be interactions across data types. RF is thus well-suited to the study of phenotypes with complex underlying etiologies, where the biological features of interest have yet to be elucidated.

While the results of this study are promising, there are questions yet to be addressed. The combined RF approach needs to be applied to a real dataset (and the results tested at the lab bench) to confirm the conclusions of the simulation study. Currently, the dataset used as the template for the simulations is being analyzed using the integrated RF approach found to be successful with these simulations. Additionally, work must be continued on modifications to RF that allow for the discovery of purely epistatic genetic models [19]. Because RF chooses only one attribute at each tree split during construction, strictly epistatic (*i.e.* absence of even miniscule main effects) attributes will not be selected. Finally, strategies for automatically translating the features selected by RF into meaningful biological hypothesis need to be developed.

5. ACKNOWLEDGMENTS

This work was supported by NIH grants AI064625, AI59694, and AI057661.

6. REFERENCES

[1] J. Lazarou, B. H. Pomeranz, and P. N. Corey, "Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies," *JAMA*, vol. 279, no. 15, pp. 1200-1205, 1998.

[2] R. A. Wilke, D. M. Reif, and J. H. Moore, "Combinatorial pharmacogenetics," *Nature Reviews Drug Discovery*, vol. 4, no. 11, pp. 911-918, 2005.

[3] J. H. Moore, "The ubiquitous nature of epistasis in determining susceptibility to common human diseases," *Human Heredity*, vol. 56, no. 1-3, pp. 73-82, 2003.

[4] J. H. Moore and S. M. Williams, "Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis," *Bioessays*, vol. 27, no. 6, pp. 637-646, 2005.

[5] C. Kooperberg and I. Ruczinski, "Identifying interacting SNPs using Monte Carlo logic regression," *Genetic Epidemiology*, vol. 28, no. 2, pp. 157-170, 2005.

[6] J. Marchini, P. Donnelly, and L. R. Cardon, "Genome-wide strategies for detecting multiple loci that influence complex diseases," *Nature Genetics*, vol. 37, no. 4, pp. 413-417, 2005.

[7] M. R. Nelson, S. L. Kardia, R. E. Ferrell, and C. F. Sing, "A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation," *Genome Res.*, vol. 11, no. 3, pp. 458-470, 2001.

[8] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *American J Human Genetics*, vol. 69, no. 1, pp. 138-147, 2001.

[9] N. Tahri-Daizadeh, D. A. Tregouet, V. Nicaud, N. Manuel, F. Cambien, and L. Tiret, "Automated detection of informative combined effects in genetic association studies of complex traits," *Genome Res.*, vol. 13, no. 8, pp. 1952-1960, 2003.

[10] L. Hood, "Systems biology: integrating technology, biology, and computation," *Mech. Ageing Dev.*, vol. 124, no. 1, pp. 9-16, 2003.

[11] D. M. Reif, B. C. White, and J. H. Moore, "Integrated analysis of genetic, genomic, and proteomic data," *Expert Reviews in Proteomics*, vol. 1, no. 1, pp. 67-75, 2004.

[12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[13] K. L. Lunetta, L. B. Hayward, J. Segal, and E. P. Van, "Screening large-scale association study data: exploiting interactions using random forests," *BMC Genet*, vol. 5, no. 1, p. 32, 2004.

[14] M. Robnik-Sikonja, "Improving random forests," *Machine Learning: Ecml 2004, Proceedings*, vol. 3201, pp. 359-370, 2004.

[15] B. A. McKinney, D. M. Reif, M. T. Rock, K. M. Edwards, S. F. Kingsmore, J. H. Moore, and J. E.

Crowe, Jr., "Cytokine expression patterns associated with systemic adverse events following smallpox immunization," *J Infect. Dis.*, vol. 194, no 4, pp. 444-453, 2006.

[16] M. T. Rock, S. M. Yoder, T. R. Talbot, K. M. Edwards, and J. E. Crowe, Jr., "Adverse events after smallpox immunizations are associated with alterations in systemic cytokine levels," *J Infect. Dis.*, vol. 189, no. 8, pp. 1401-1410, 2004.

[17] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. New York: Chapman & Hall, 1984.

[18] M. A. Province, W. D. Shannon, and D. C. Rao, "Classification methods for confronting heterogeneity," *Adv Genetics*, vol. 42, pp. 273-286, 2001.

[19] A. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, and E. P. Van, "Identifying SNPs predictive of phenotype using random forests," *Genet Epidemiology*, vol. 28, no. 2, pp. 171-182, 2005.

[20] B. A. McKinney, D. M. Reif, M. D. Ritchie, and J. H. Moore, "Machine Learning for Detecting Gene-Gene Interactions: A Review," *Appl. Bioinformatics*, vol. 5, no. 2, pp. 77-88, 2006.

[21] R. Culverhouse, B. K. Suarez, J. Lin, and T. Reich, "A perspective on epistasis: limits of models displaying no main effect," *American J Human Genetics*, vol. 70, no. 2, pp. 461-471, 2002.

[22] R. Ihaka and R. Gentleman, "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics*, vol. 5, pp. 299-314, 1996.

[23] R Development Core Team, "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, 2006.

[24] L. Breiman and A. Cutler, "Random Forests," www.stat.berkeley.edu/users/breiman/RandomForests, 2004.

View publication stats