BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 18 2008, pages 2010–2014
doi:10.1093/bioinformatics/btn356

Gene expression

Enriched random forests

Dhammika Amaratunga[1,*], Javier Cabrera[2] and Yung-Seop Lee[3]

[1]Department of Nonclinical Biostatistics, Johnson & Johnson PRD LLC, Raritan, NJ 08869, [2]Department of Statistics, Rutgers University, 110 Frelinghuysen Ave, Piscataway, NJ 08854, USA and [3]Department of Statistics, Dongguk University, Seoul, South Korea

Received and revised on July 1, 2008; accepted on July 10, 2008

Advance Access publication July 22, 2008

Associate Editor: David Rocke

ABSTRACT

Although the random forest classification procedure works well in datasets with many features, when the number of features is huge and the percentage of truly informative features is small, such as with DNA microarray data, its performance tends to decline significantly. In such instances, the procedure can be improved by reducing the contribution of trees whose nodes are populated by non-informative features. To some extent, this can be achieved by prefiltering, but we propose a novel, yet simple, adjustment that has demonstrably superior performance: choose the eligible subsets at each node by weighted random sampling instead of simple random sampling, with the weights tilted in favor of the informative features. This results in an 'enriched random forest'. We illustrate the superior performance of this procedure in several actual microarray datasets.

Contact: damaratu@prdus.jnj.com

1 INTRODUCTION

The random forest is a popular classification technique whose classifier is an ensemble of classification trees. It has developed an excellent reputation amongst the statistics and machine learning communities as a versatile method that produces accurate classifiers for many types of data. It is considered particularly well suited to situations characterized by a large number of features, a circumstance that is becoming more prevalent as the ability to collect and store vast amounts of data becomes easier and increasingly common. In such instances, classical classification approaches tend to become overwhelmed by the number of features and fail. Yet random forest continues to do well. For instance, with DNA microarray data, work by Dudoit *et al*. (2002), Lee *et al*. (2005) and Díaz-Uriarte and de Andrés (2006) shows that random forest outperforms most of the other classification techniques.

However, when, in addition to having a large number of features, the proportion of truly informative features is small, its performance too tends to decline. An example that illustrates this point, in fact the case that motivated us to look into this problem, is an experiment conducted to study whether mice whose Slc17A5 gene had been knocked out could be distinguished from wild-type mice at the gene expression level (Moechars *et al*., 2005; Raghavan *et al*., 2007). Gene expression measurements were taken on newborn (0-day-old) mice as well as on 18-day-old mice using Affymetrix Mouse430_2 GeneChips. The day 0 time point preceded the occurrence of any obvious phenotypic variations in the knockout mice but subtle effects would already have begun at the cellular level. By day 18 phenotypic variations in the knockout mice are evident with observable morphological alterations such as defects in myelination. Thus separation of the 18-day-old mice is straightforward both physiologically and with gene expression data; in fact random forest does this with an out-of-bag error rate of $<10\%$. On the other hand, the newborn mice are a challenge to separate: impossible physiologically and difficult even with gene expression data; the out-of-bag error rate for random forest is over 50%.

We shall show that, in situations like this, the basic random forest procedure can be considerably enhanced by reducing the contribution of trees whose nodes are populated by less informative features, resulting in a procedure which we refer to as the enriched random forest procedure. Doing this shrinks the out-of-bag error rates for both the Slc17A5 datasets to 0%, a tremendous achievement particularly with the newborn mice. Our ability to detect a gene expression signal at day 0, despite the lack of any obvious external signs, connotes the presence of early genomic effects and is important biological information.

In the following sections, we will describe random forest and the novel modifications we are proposing and then we will assess the performance of the standard and enriched procedures via several microarray datasets. We consider only microarray data in this article, although the methods are also potentially applicable to other megavariate situations such as mass spectrometry and molecular imaging as well.

2 METHODS

2.1 Random forest

We begin with a brief outline of the random forest algorithm; Breiman (2001) and Breiman and Cutler (2003) provide further details.

Given a training set X comprised of N cases, which belong to two classes, and G features, a classification tree can be constructed as follows. First, a feature x and a threshold t that splits X into two subsets that are maximally distinct according to a specified criterion are selected from all features of x and all possible values of t. The training set is then split into the two buckets X_L and X_R depending on whether or not $x < t$. This procedure is repeated with each of X_L and X_R using another (x, t) combination. This process is repeated until no further splitting is possible.

In a random forest, a tree, rather than being trained on the entirety of the training set, is trained only on a sample of N cases drawn at random with replacement from the complete set of N cases. This is the bagging

[*] To whom correspondence should be addressed.

(Breiman, 1996) step and the selected samples are called the in-bag cases; the rest are set aside as out-of-bag cases. Additionally, when determining which feature to split on at each node, only a subset of g of the G features (usually $g = G^{1/2}$) are considered eligible; this subset is drawn at random without replacement independently for each node from the complete set of G features.

A random forest is an ensemble of some number R of such trees, where each tree is called a base classifier. Classes are assigned to test cases by majority vote: when given a test case, each tree assigns it a class according to its classifier; this information is collated and overall the forest assigns it the most frequent class. The out-of-bag cases in any tree can be regarded as test cases for that tree as they were not used to build it and thus they can be used to assess the performance of the forest as a whole; this is done via the out-of-bag error rate, which is the proportion of times an out-of-bag case is misclassified.

Most classification procedures when faced with a situation in which there are a large number of features exhibit a tendency to overfit. However, a major advantage of random forests is that they are able to keep the likelihood of overfitting low by using different subsets of the training data and different subsets of features for training the different base classifiers. Thus only patterns truly present in the data would be detected consistently by a majority of the base classifiers and the majority votes turn out to be good indicators of class.

In general, for the random forest classifier to be effective, each base classifier must have reasonably good classification performance and the trees must be diverse and only weakly correlated. The first of these, individual performance, is obtained by using strong performing tree classifiers and the second, diversity, is obtained by randomly choosing cases on which to train each tree and by randomly choosing attributes at each node of each tree.

2.2 Enriched random forest

When the number of possible features is huge and the percentage of truly informative features is small, a problem arises. The performance of the base classifiers degrades. This is because, if simple random sampling is used for selecting the subset of g eligible features at each node, almost all these subsets are likely to contain a preponderance of non-informative features. Consider a situation with G features, of which only H are informative. Then, if at any node g features are selected by resampling randomly with equal weights, the probability distribution of the number of informative features selected is binomial with g trials and probability $\pi = H/G$, so that the mean number of informative features selected at each iteration is $\mu = \pi g$. Since π is typically tiny, so will μ be. For example, if H= 100, G=10 000 and $g = G^{1/2}$= 100, the resulting μ is only one informative feature per node. The base classifiers built using such nodes will have low accuracies and overall, the performance of the ensemble will suffer.

This can be remedied by using weighted random sampling instead of simple random sampling [an analogous issue arising in unsupervised classification is addressed similarly by Amaratunga *et al*. (2008)]. By tilting the random sampling of genes so that less informative genes are less likely to get selected, the odds of trees containing more informative features being included in the forest increases. Consequently, the resultant random forest, which we call the 'enriched random forest' (ERF), will be comprised of a higher number of better base classifiers, resulting in a better fit. Due to the enormous choice of features available, the diversity of the ensemble is not compromised and can be controlled to be more diverse than prefiltering. The value of doing the weighting is amply demonstrated in the performance improvement over both non-filtering and prefiltering reported in Section 3.

Weighting can be done by scoring each gene based on its ability to separate the groups, e.g. via a t-test, and using these scores to assign weights, $\{w_i\}$, so that the genes that most separate the groups are the only ones assigned high weights. Once the weights have been determined, the algorithm as described earlier is run with the only modification being that when, at any node, the subset of g eligible features is selected, it is selected from the G features using weighted random sampling with weights $\{w_i\}$ rather than simple random sampling.

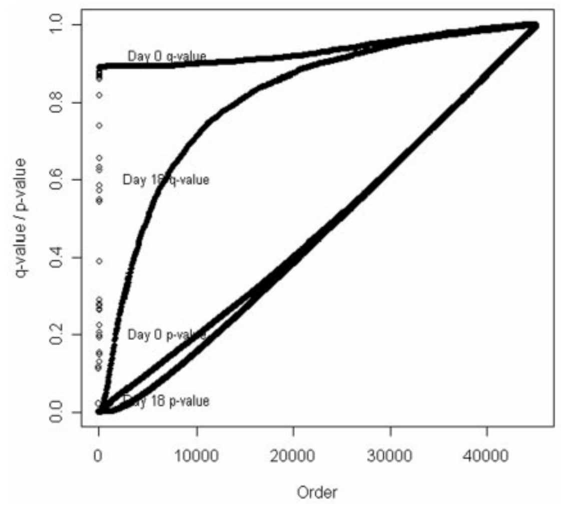

Fig. 1. Plot of p-values and q-values for the two Slc17A5 datasets.

2.3 Weighting the genes

The key to the modified algorithm is to score each feature based on how well it separates the two groups. Such a score can be generated by testing each feature for a group mean effect using a two-sample t-test and calculating the p-value, small p-values indicating greater separation and large p-values indicating less separation. However, to weight using the t-test p-values themselves would be inappropriate due to (i) the multiplicity of tests being performed and (ii) the small sample sizes typical of microarray experiments.

To adjust for (i), we base the weights on q-values (Storey and Tibshirani, 2003), which are calculated from the p-values as: q_i= $\min_{k\geq 1}\{\min((G/k)p_{(k)},1)\}$, where $p_{(i)}$ and $q_{(i)}$ are the p-value and q-value associated with the feature with the i-th smallest p-value. The q-values provide false discovery rate (FDR)-adjusted measures of significance for the features and are in the same order as the p-values.

The q-value distributions for the two Slc17A5 datasets (Fig. 1) display the difference between them. On day 0, very few genes show differential expression, indicating that most of a random forest's base classifiers would be comprised almost entirely of genes with little separability information and, as a result, it would perform~~s~~ poorly. On day 18, with many genes showing differential expression, random forest would have no problem. On the other hand, the two p-value distributions are similar to each other and notably the day 0 p-value distribution fails to highlight the few genes that separate the groups on that day. Thus, assigning weights inversely proportional to the q-values is likely to emphasize the separating genes and strengthen the likelihood of an enriched procedure detecting the separation in both datasets but assigning weights inversely proportional to the p-values is unlikely to be productive for the day 0 data.

The use of q-values rather than p-values also helps to lessen the likelihood of overfitting in situations with no separation of the data into groups. In that case, $p \sim$ unif(0,1) approximately, implying that $p_{(k)} \cong k/G$ and $q_{(k)} \cong 1$ for all k. If p-value-based weights were used, some genes would by chance have small p-values and would wrongly be assigned high weights and as a result ERF could mistakenly imply a separation. If q-value-based weights were used, all genes would be assigned equal weights and ERF should not find a separation.

Thus it is reasonable to use the q-values to assign weights: $w_i{}' = (1/q_i) - 1$. Based on this weighting, features with less separability (which will have both $p_i \cong 1$ and $q_i \cong 1$) will get almost zero weight and features with high separability will get large weights as desired. We make two small modifications to these potential weights.

First, to prevent a large fraction of features getting essentially zero weight and reducing too much the diversity of features across the eligible subsets, we threshold the lower weights to a small positive value, a_{min} This is particularly important in instances where there is no separation in the data. In such a situation it is possible that by chance a few genes could have a few non-unit q-values; in that case, if all the unit q-values are assigned zero weight, then the few non-unit q-values will receive large weights and possibly imply a faux separation. Alternatively it is also possible when there is no separation for all the q-values to be 1, which would lead to an algorithmic impasse if there was no positive threshold.

Second, to prevent a small number of features with tiny p-values from receiving huge weights, which would also drastically reduce the diversity of the ensemble, we threshold the high weights to a_{max}.

Thus the weights assigned are: w_i = median(a_{min}, $w_i{}'$,a_{max}). We set a_{min} = 0.01 and a_{max} = 999, which gives a suitably wide range of variation (of order 10^5) across weights.

To adjust for (ii), we use Ct or Conditional t (Amaratunga and Cabrera, 2007) in place of the usual t-test. The usual t-test is the simplest way to assess separation for a feature, but since microarray experiments typically have small sample sizes, the t-test has low power and thus low discriminatory ability. Therefore, analysis methods that borrow strength across features, such as Ct or limma (Smyth, 2004), are likely to generate a better ranking of features.

Since error rates could be underestimated if the weights are calculated outside the bagging loop (i.e. calculated just once based on all the samples), they are calculated inside it, i.e. they are determined separately for each tree based on only the in-bag samples, so that they are independent of the out-of-bag samples. However, this does increase the computational burden and also renders the weights somewhat less well determined than if they had been calculated outside the loop using all the samples.

In procedure ERF-CV, a variant of ERF, balanced leave-out-one cross-validation is used instead of bagging to lighten the computational load and to decrease its loss of sample size when determining weights. Let $J=R/N$. In ERF-CV, in J of the R trees, Case 1 is set aside as the out-of-bag test set, the weights are calculated based on the $N-1$ in-bag cases and a tree is derived based on these in-bag cases and out-of-bag prediction is done on Case 1. This is repeated with each of the other cases. Less computation is required for ERF-CV than for ERF since weights are calculated only N times rather than R times.

3 RESULTS

The best way to assess the performance of the enriched random forest procedure would be to apply it to a dataset which we know from context has a subtle signal. After all, it is in this situation that ERF works best and is superior to RF; if the signal were strong or non-existent, both ERF and RF would produce essentially the same result. Since the context needs to be known, such datasets are hard to come by in the public domain despite the ready availability of microarray datasets on the internet. Thus, we will use the Slc17A5 Day 0 dataset as the primary vehicle for our evaluation of ERF. The Slc17A5 Day 18 dataset, which has an unequivocal separation of classes, will be used to assess the performance of ERF when there is a strong signal in the data. To study the other extreme, two datasets that have no signal were created artificially by scrambling the samples of the two Slc17A5 datasets. These datasets will be used to verify that the methods are not overfitting. If the weighting is not done carefully, it is possible to find spurious classifications in datasets that have no true separation. In addition, ERF was run on eight public domain DNA microarray datasets whose sample classifications were known. The 10 datasets are listed in Table 1. All data were quality checked, log transformed and quantile normalized as described in Amaratunga and Cabrera (2004).

The out-of-bag error rates are shown in Table 2. Good classification methods should show low out-of-bag error rates for the eight original datasets and high out-of-bag error rates for the two scrambled datasets. Besides standard random forest and enriched random forest with t- and Ct-based weights, since prefiltering is sometimes used to improve classifier performance in microarray settings (Li and Yang, 2002), we also ran random forest after filtering out any gene that was not significant at the 25% level. Another way to improve performance in microarray settings is to increase g, so we also ran random forest with $g=G^{3/4}$. Incidentally, there is nothing special about the 25% level or $G^{3/4}$ but other values were tried and none stood out as consistently superior. In addition, we ran the VarSelRF procedure (Díaz-Uriarte and de Andrés, 2006) as a benchmark for a recent random forest procedure designed for microarray data analysis.

Where there is true separation between groups, it can be seen that the enriched random forest performs consistently equally or better than the standard random forest. ERF's best performance over RF is when the number of genes is huge (over 40 000), such as with the two Slc17A5 datasets and the Prostate Cell Type dataset. When RF detects a clear classification as with the Human Lymph Node Sinus and Breast Cancer datasets, so does ERF. When there is no signal, such as with the scrambled datasets, ERF, like RF, correctly finds none.

Table 1. Datasets used in the performance assessment

Dataset name	No. of genes	Samples	Reference
Slc17A5 Day 0	45 101	Wild type (6) versus knockout (6)	Raghavan *et al*. (2007)
Slc17A5 Day 18	45 101	Wild type (6) versus knockout (6)	Raghavan *et al*. (2007)
Adenocarcinoma	13 432	Metastasis (12) versus tumor (64)	Ramaswamy *et al*. (2002)
Astrocytoma	12 625	Low grade (8) versus high grade (6)	MacDonald (2001)
Breast cancer	15 926	Normal (11) versus patients (24)	Chan *et al*. (2005)
Diabetes	22 283	Normal (17) versus diabetes (18)	Mootha *et al*. (2003)
Epilepsy	31 099	Control (6) versus phenytoin (7)	Wilson *et al*. (2005)
HIV Encephalitis	12 625	Reference (12) versus encephalitis (16)	Masiliah *et al*. (2004)
Human Lymph Node Sinus	22 283	Tonsils (10) versus lymph node (10)	Martens *et al*. (2006)
Prostate Cell Type	54 675	Endothelial (5) versus stromal (5)	Oudes *et al*. (2006)

Table 2. Out-of-bag error rates for random forest (RF), random forest with *p*-value based filtering (RF(*p*)), random forest with *g*-based filtering (RF(*g*)), the VarSelRF procedure (VRF), enriched random forest with *t*-based weights (ERF(*t*) and ERF-CV(*t*)) and enriched random forest with Ct-based weights (ERF(Ct) and ERF-CV(Ct))

	RF	RF(*p*)	RF(*g*)	VRF (Boot)	VRF (CV)	ERF (*t*)	ERF (Ct)	ERF-CV (*t*)	ERF-CV (Ct)
Slc17A5 Day 0	0.583	0.583	0.250	0.527	0.543	0.167	0.000	0.083	0.000
Slc17A5 Day 18	0.083	0.083	0.083	0.260	0.316	0.000	0.000	0.000	0.000
Slc17A5 Day 0 (scrambled)	*0.750*	*0.750*	*0.333*	*0.535*	*0.556*	*0.833*	*0.667*	*0.750*	*0.667*
Slc17A5 Day 18 (scrambled)	*0.583*	*0.667*	*0.583*	*0.494*	*0.496*	*0.667*	*0.667*	*0.417*	*0.583*
Adenocarcinoma	0.132	0.145	0.145	0.211	0.229	0.145	0.145	0.145	0.158
Astrocytoma	0.214	0.071	0.000	0.229	0.284	0.000	0.071	0.071	0.071
Breast Cancer	0.029	0.029	0.029	0.041	0.061	0.029	0.029	0.029	0.000
Diabetes	0.543	0.514	0.286	0.387	0.422	0.457	0.543	0.486	0.571
Epilepsy	0.154	0.077	0.000	0.306	0.357	0.154	0.154	0.077	0.154
HIV Encephalitis	0.357	0.429	0.357	0.283	0.335	0.250	0.250	0.250	0.286
Human Lymph Node Sinus	0.000	0.000	0.000	0.056	0.083	0.000	0.000	0.000	0.000
Prostate Cell Type	0.200	0.200	0.300	0.295	0.348	0.100	0.100	0.100	0.100

R=1000 trees were used throughout. The out-of-bag error rates for the scrambled datasets are italicised.

When RF fails to detect an existing separation, as with the Slc17A5 Day 0 dataset and less starkly the Astrocytoma dataset, prefiltering is only slightly helpful and sequential filtering as in VarSelRF is of no help at all. Interestingly, increasing *g* turns out to be a more effective option. However, in practice it is unclear as to how much filtering would be useful or what value of *g* would be the best for any given dataset. Weighted sampling as in ERF, on the other hand, provides a softer filter and the process of combining multiple classifiers allows the procedure to adapt to the situation.

Both *t*-based and Ct-based ERF procedures gave good results. In experiments with very small sample sizes, such as the Slc17A5 experiments, Ct gives lower error rates that we conjecture is due to the increase in power induced by borrowing strength across genes. An added advantage of the bootstrap-based Ct procedure is that it is distribution-free. In contrast, for the Day 0 and Day 18 Slc17A5 datasets, the Wilcoxon test, which is also distribution-free but does not borrow strength across genes, gives non-competitive ERF-CV error rates of 0.583 and 0.083, respectively. Limma, which borrows strength across genes but is not distribution-free, gives ERF-CV rates of 0.083 and 0.000, respectively, that are close to the Ct-based ERF-CV error rates.

Overall, both ERF and ERF-CV performed well and were improvements over standard random forest. By and large, the ERF and ERF-CV error rates were similar to each other. Thus, ERF-CV may be more useful in practice since it is less computationally intensive and less prone to small sample size problems.

The weights we use here are also useful for data displays. Principal components analysis plots (Fig. 2) for the two Slc17A5 datasets show no separation between the two groups at day 0 when equal weights are used, whereas with the Ct *q*-value-based weights, the separation is clear in the first principal component itself. For the day 18 data, the separation is clear without weights but far clearer with weights. Here, the weights were incorporated into the calculation of the covariance matrix by making the variance of each gene equal to the corresponding weight assigned to the gene.

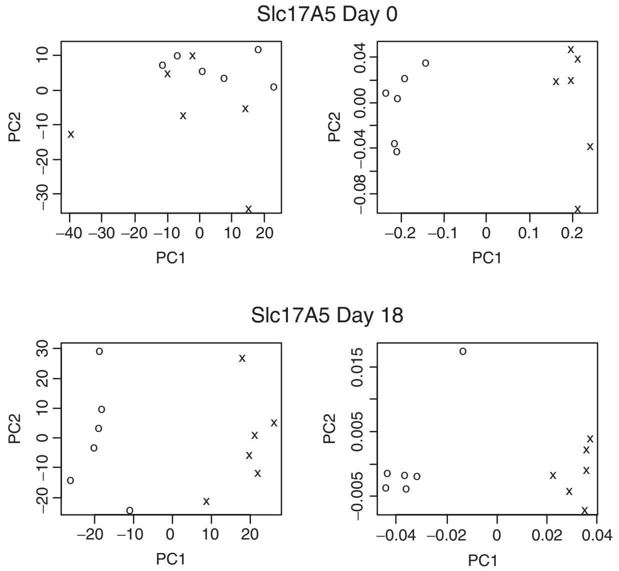

Fig. 2. Principal components analysis plots for the two Slc17A5 datasets (o = wild-type, x = knockout) using equal weights (left) and Ct-based weights (right).

4 DISCUSSION

As technology advances, microarrays capable of interrogating an increasingly large number of transcripts are being deployed. For instance, currently Affymetrix's popular Human Genome array set contains in excess of 40 000 probesets, double that from just a few years ago. Experiments that employ such arrays will generate megavariate data.

It is hard to detect subtle signals in megavariate data. Yet, this is an important practical problem as it is in these situations where classification *per se* is useful. If the signal can be detected easily, then the interest is more in trying to find the genes, gene sets (e.g. gene sets defined by GO terms) or signature associated with the signal and this is best accomplished using a combination of feature selection methods and biological information [as in Raghavan *et al*. (2007) for the Slc17A5 datasets]. Using classification alone for feature selection carries overfitting risk due to the huge number of features and the emphasis on the given structure (Strobl *et al*., 2007).

When the interest is in classification itself, we have, in this article, offered a novel solution: a simple enhancement to the familiar random forest procedure that greatly improves its performance in situations in which the number of features is huge but the proportion of informative features is small. R code for ERF is available at the websites: http://www.geocities.com/damaratung/ and http://www.rci.rutgers.edu/∼cabrera/DNAMR/.

We have only discussed ERF in the two-group context here, but ERF can be extended to the case of multiple (i.e. more than two) groups. However, here the complexity grows as often the features that separate any two groups could differ substantially from the features that separate any two other groups. Thus, this situation calls for a more complex solution possibly involving collation of multiple pairwise analyses that is beyond the scope of this article.

In addition to supervised classification by random forest, the idea of using weighted random sampling of features instead of simple random sampling has been shown to be effective in the unsupervised classification problem as well (Amaratunga *et al*., 2008). Hence, we conjecture that this idea could be incorporated into other ensemble and machine learning techniques such as linear discriminant analysis, logistic regression and support vector machines. We plan to continue developing this work in that direction.

ACKNOWLEDGEMENTS

The authors would like to express their thanks to Dieder Moechars and Hinrich Göhlmann of Johnson & Johnson PRD LLC for making the Slc17A5 data available to us, to the investigators who make their data publicly available via ArrayExpress and GEO, to James Colaianne, also of Johnson & Johnson PRD LLC, for his support during this project, and to the referees and the Associate Editor for their insightful suggestions that greatly improved the article.

Funding: J.C. was funded in part by NSF grant DBI-0629346. Y.-S.L. was supported by the research program of Dongguk University.

Conflict of Interest: none declared.

REFERENCES

Amaratunga,D. and Cabrera,J. (2004) *Exploration and Analysis of DNA Microarray and Protein Array Data*. John Wiley, New York.

Amaratunga,D. and Cabrera,J. (2007) A conditional t suite of tests for identifying differentially expressed genes in a DNA microarray experiment with little replication. *Stat. Biopharmaceut. Res.* Available at www.amstat.org/publications/sbr/index.cfm?fuseaction=amaratunga

Amaratunga,D. *et al*. (2008) Microarray learning with ABC. *Biostatistics*, **9**, 128–136.

Breiman,L. (1996) Bagging predictors. *Mach. Learn.*, **24**, 123–140.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Breiman,L. and Cutler,A. (2003) Random forests manual (version 4.0). *Technical Report of the University of California*, *Berkeley*, *Department of Statistics*.

Chan,M.M. *et al*. (2005) Gene expression profiling of NMU-induced rat mammary tumors: cross species comparison with human breast cancer. *Carcinogenesis*, **26**, 1343–1353.

Díaz-Uriarte,R. and de Andrés,S.A. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3.

Dudoit,S. *et al*. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.

Lee,J.W. *et al*. (2005) An extensive evaluation of recent classification tools applied to microarray data. *Comput. Stat. Data Anal.*, **48**, 869–885.

Li,W. and Yang,Y. (2002) How many genes are needed for a discriminant microarray data analysis? In Lin,S.M. and Johnson,K.F. (eds) *Methods of Microarray Data Analysis*. Kluwer, Boston, pp. 137–150.

MacDonald,T. (2001) Human glioblastoma. Available at http://pepr.cnmcresearch. org/browse.do?action=list_prj_exp&projectId=65

Martens,J.H. *et al*. (2006) Differential expression of a gene signature for scavenger/lectin receptors by endothelial cells and macrophages in human lymph node sinuses, the primary sites of regional metastasis. *J. Pathol.*, **208**, 574–589.

Masiliah,E. *et al*. (2004) Patterns of gene dysregulation in the frontal cortex of patients with HIV encephalitis. *J. Neuroimmunol.*, **157**, 163–175.

Moechars,D. *et al*. (2005) Sialin-deficient mice: a novel animal model for infantile free sialic acid storage disease (ISSD). In *Society for Neuroscience 35th Annual Meeting*. Washington, DC, USA.

Mootha,V.K. *et al*. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.

Oudes,A. *et al*. (2006) Transcriptomes of human prostate cells. *BMC Genomics*, **7**, 92.

Raghavan,N. *et al*. (2007) The high-level similarity of some disparate gene expression measures. *Bioinformatics*, **23**, 3032–3038.

Ramaswamy,S. *et al*. (2002) A molecular signature of metastasis in primary solid tumors. *Nat. Genet.*, **33**, 49–54.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article 3.

Storey,J.D. and Tibshirani,R. (2007) Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci.*, **100**, 9440–9445.

Strobl,C. *et al*. (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, **8**, 25.

Wilson,D.N. *et al*. (2005) Microarray analysis of postictal transcriptional regulation of neuropeptides. *J. Mol. Neurosci.*, **25**, 285–298.