BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 14 2010, pages 1752–1758
doi:10.1093/bioinformatics/btq257

Data and text mining

Advance Access publication May 26, 2010

On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data

Daniel F. Schwarz, Inke R. König* and Andreas Ziegler*

Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Maria-Goeppert-Strasse 1, 23562 Lübeck, Germany

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Genome-wide association (GWA) studies have proven to be a successful approach for helping unravel the genetic basis of complex genetic diseases. However, the identified associations are not well suited for disease prediction, and only a modest portion of the heritability can be explained for most diseases, such as Type 2 diabetes or Crohn's disease. This may partly be due to the low power of standard statistical approaches to detect gene–gene and gene–environment interactions when small marginal effects are present. A promising alternative is Random Forests, which have already been successfully applied in candidate gene analyses. Important single nucleotide polymorphisms are detected by permutation importance measures. To this day, the application to GWA data was highly cumbersome with existing implementations because of the high computational burden.

Results: Here, we present the new freely available software package Random Jungle (RJ), which facilitates the rapid analysis of GWA data. The program yields valid results and computes up to 159 times faster than the fastest alternative implementation, while still maintaining all options of other programs. Specifically, it offers the different permutation importance measures available. It includes new options such as the backward elimination method. We illustrate the application of RJ to a GWA of Crohn's disease. The most important single nucleotide polymorphisms (SNPs) validate recent findings in the literature and reveal potential interactions.

Availability: The RJ software package is freely available at http://www.randomjungle.org

Contact: inke.koenig@imbs.uni-luebeck.de; ziegler@imbs.uni-luebeck.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 9, 2009; revised on April 16, 2010; accepted on May 16, 2010

1 INTRODUCTION

Genome-wide association (GWA) studies have become a standard approach for helping unravel the genetic basis of complex genetic diseases. The recent successes are tremendous, and a series of new loci have been identified using single marker analyses (McCarthy *et al*., 2008; Samani *et al*., 2007; Wellcome Trust Case Control Consortium, 2007). Unfortunately, only a small portion of the heritability was explained by corresponding single nucleotide polymorphisms (SNPs) for most diseases such as Type 2 diabetes (6%) or Crohn's disease (20%) (Manolio *et al*., 2009). Furthermore, SNPs identified by GWA studies for various diseases make poor classifiers (Jakobsdottir *et al*., 2009).

For overcoming such drawbacks and recognizing the complexity of the underlying biology, further mechanisms such as gene–gene interaction need to be taken into account (Moore *et al*., 2010). However, the discovery of gene–gene interactions using GWA studies remains challenging with traditional statistical approaches (Cordell, 2009; Moore *et al*., 2010). Given genotype data at different loci, an exhaustive search of interactions between all loci is the obvious way of testing interactions. Testing all two-locus interactions is computationally feasible although time demanding (Marchini *et al*., 2005). However, an exhaustive search of higher order interactions is computationally impractical because the number of tests increases exponentially with the order of interaction (Cordell, 2009).

One approach to deal with such large numbers of SNPs is to first perform univariate tests on each SNP, discard SNPs with high P-values and apply interaction methods, e.g. within logistic regressions, to SNP subsets afterwards (Hoh *et al*., 2000; Marchini *et al*., 2005). Unfortunately, such approaches may result in low power for SNP–SNP interactions with very small marginal effects.

Another concern is genetic heterogeneity, i.e. different subsets of genes affect the same disease, and traditional statistical methods show limitations when genetic heterogeneity is present (Province *et al*., 2001).

A promising alternative is Random Forests (RFs; Breiman, 2001). RF was applied successfully to genetic data in various studies (Bureau *et al*., 2005; Chang *et al*., 2008; Jiang *et al*., 2009; McKinney *et al*., 2009; Sun *et al*., 2007), and it is anticipated that RF will help to detect gene–gene interactions in genome-wide data (Moore *et al*., 2010). It has been shown that RF can substantially be more efficient than standard statistical methods in ranking the true disease-associated SNPs in order to detect SNP–SNP interaction (Lunetta *et al*., 2004). The method is able to detect SNPs with small effects and to deal with genetic heterogeneity because separate models are automatically fit to subsets of data defined by early splits in the tree (Lunetta *et al*., 2004; Province *et al*., 2001). In addition, RF is able to handle SNPs that are associated in a non-linear fashion.

The RF method is a specific data mining method. In data mining, in general, algorithms attempt to identify an unknown concept based on randomly chosen examples of the collected data. The aim is to find a prediction rule that correctly classifies new instances of the

*To whom correspondence should be addressed.

concept (Breiman, 2001). Thus, RF makes fewer assumptions about the functional form of the model, as required by statistical tests (McKinney *et al*., 2006).

A grown tree in a forest is often graphically represented by an upside down tree. Multiple paths lead through the tree from the root to different leaves via various nodes. Each node corresponds to a specific predictor variable. Thus, a path is a sequence of predictor variables (for details, see Section 2.1 and König *et al*., 2008). Such a predictor variable sequence includes potential interactions between them in terms of hierarchical dependencies (Cordell, 2009; Moore *et al*., 2010). Thus, the RF method allows for interactions between SNPs.

RF yields a classification result and a measure of the importance for each variable. Variable importance (VI) quantifies the impact of a SNP in predicting the response and may reflect a causal effect. In turn, it can be used to select the relevant SNPs from a GWA study (Ziegler *et al*., 2007).

Although appealing, RF has rarely been applied on the genome-wide level. In analogy to standard statistical approaches, this is due to the computational intensity and memory requirements (Zhang *et al*., 2009; Ziegler *et al*., 2007). The original RF implementation, termed RF in Fortran, by Breiman and Cutler (2004) designed to analyze low-dimensional data, i.e. a low number of SNPs, with a large number of observations, e.g. 100 SNPs and 10 000 observations. It has been successfully used, e.g. by Bureau *et al*. (2005) in a candidate gene case–control study involving 42 SNPs. However, it is computationally and memory inefficient so that not more than 10 000 of SNPs can be analyzed on a standard machine within reasonable time and memory usage (Ziegler *et al*., 2007). Furthermore, the code is not user-friendly because the program has to be modified and compiled anew, whenever a new dataset is used.

An alternative implementation is the randomForest package for the programming language R (R Development Core Team, 2009) by Liaw and Wiener (2002). It is user-friendly, and it has been often used in applications (Ziegler *et al*., 2007). The source code of the package randomForest consists of R, C and Fortran source code. Elementary subroutines were left in Fortran code. However, the same computational and memory limitations apply as to RF in Fortran.

One approach to overcome the memory issue is to split up the GWA data into small chunks, which are subsequently analyzed separately (Jiang *et al*., 2009; Schwarz *et al*., 2007). The results of all processed chunks are finally combined. Through this, main effects are detected, but one may fail to discover some important interaction effects due to data separation. Thus, these approaches do not overcome the restrictions in detecting complex interactions.

An alternative has recently been presented by Zhang *et al*. (2009). In their package Willows, they compress the GWA data internally and subsequently apply RF. However, it is slow for large values of the mtry parameter (see Section 2) as recommended for datasets with many noise variables such as GWA study data (Breiman and Cutler, 2004; Liaw and Wiener, 2002). Tuning mtry for optimizing the performance of the forest is also strongly recommended (Breiman and Cutler, 2004), which can hardly be done with Willows. Using this program can be computationally intensive, thus time demanding.

Here, a novel software package called Random Jungle (RJ) is presented, which has been specifically tailored for the large-scale analysis of GWA studies. This computational and memory efficient implementation of RF is able to analyze hundreds and thousands of samples and SNPs.

In the following, we first briefly introduce the RF methodology, including the growing procedure and essential features. Next, we describe the estimation of various VI measures. Specifically, we show differences between importance scores of randomForest and RF in Fortran. After these theoretical considerations, we describe the RJ software and demonstrate its superior computational performance when compared with other implementations. Finally, we illustrate its use with data from a GWA study on Crohn's disease.

2 METHODS

2.1 Random forests

RFs is an ensemble consisting in multiple classification and regression trees (CART) that are grown using a bootstrap sample of given data and without pruning. In general, an ensemble is a group of classifiers in which the classifier is only required to perform slightly better than random guessing or coin flipping. This property is fulfilled by many base classifiers, such as CART (Breiman, 1996; Schapire, 1990). With a CART as base classifier, a sample is classified by taking the majority vote over all tree classifiers in a forest (Breiman, 2001). RF has been shown to provide good accuracy, robustness to noise, internal estimation of error, stable classifiers and VI (Breiman, 2001; Breiman and Cutler, 2004; Meng *et al*., 2009). The RF procedure takes the following steps (Breiman, 2001):

(1) Consider a dataset X, termed training data, consisting of one response variable and many predictor variables from N samples. The total count of predictor variables is M, with M being substantially larger than N.

(2) A bootstrap sample X^* consisting of N samples is drawn with replacement from the original training data X. On average, one-third of all samples are left out due to the bootstrapping process. These samples are called 'out-of-bag' (OOB) data $X \backslash X^*$.

(3) A CART t is grown using the bootstrap dataset X^*. The CART is constructed by recursively splitting data into distinct subsets, so that one parent node leads to two child nodes. For splitting data, an appropriate split rule has to be selected so that the subsets of each child node are purer than the subset of corresponding parent node. The goodness of the split is defined to be the decrease in impurity as follows: $\Delta i = i_{\text{parent}} - (p_{\text{left}} \cdot i_{\text{left}} + p_{\text{right}} \cdot i_{\text{right}})$. The proportion of samples in left and right nodes is given by p_{left} and p_{right}, respectively. The measure of impurity i_{parent}, i_{left} and i_{right} of parent node, left and right child node is determined by the Gini index, i.e. $i = 1 - \Sigma_j p(j)^2$, where is the proportion of samples that are labeled with class j in that node. At each node, a random subset of all predictor variables is chosen without replacement to determine the best split. The size of the subset is given by the parameter mtry. Although different variables might be selected at each node to be tested, the number mtry is held constant during the procedure, and the default setting is mtry $= \left\lceil \sqrt{M} \right\rceil$, where $\lceil \cdot \rceil$ denotes the next larger integer.

(4) The tree t is grown to its largest extent, and no pruning proceeds. The final nodes are called terminal nodes.

(5) Steps 1 to 4 are repeated to grow a specific number of trees, and, for classification, the majority vote over all trees in the resulting forest is used.

(6) Finally, the OOB error fraction is calculated by classifying each sample of the OOB. Each observation is predicted by the trees for which it is an OOB observation. The prediction accuracy of the classifier is estimated by subtracting the OOB error fraction from its maximum, which is one. The prediction accuracy estimation method is a suitable surrogate for cross-validation (Breiman, 2001).

A special feature of RF is the calculation of proximities between samples. For this, after a tree is grown, every subject is classified by each tree. Then, each pair of subjects is compared with regard to its final stopping point. That is, if they are assigned to the same terminal node in a single tree of the forest, the proximity between them is increased by one. The proximity matrix is useful, e.g. for replacing missing data, imputing data, identifying outliers and finding class representative samples called prototypes.

RF can be turned into an unsupervised learning method. To initialize the process, the original dataset is considered as Class 1. A new synthetic dataset of the same number of samples and predictor variables is created and labeled as Class 2. This synthetic data is created by sampling at random without replacement from the univariate distributions of the original data. The original and the synthetic data are merged. The resulting artificial two-class dataset is analyzed by RF in order to produce sample proximities as described above. The 2D multidimensional scaling (MDS) technique (Cox and Cox, 2001) is subsequently applied to the proximity matrix. The method yields a 2D graphical representation of the underlying sample structure. To identify clusters in the sample structure, the graphical representation has to be investigated by standard clustering techniques, such as k-means clustering (Macqueen, 1967).

A further feature is the computation of sample margins. A sample margin is the difference of proportional votes for the correct class and maximum proportional votes of remainder classes. Sample margins are defined between 1 and −1. A high positive sample margin means a coherent and correct classification.

The standard RF methodology can be extended by a flexible backward elimination procedure. The procedure identifies small sets of variables that can achieve good predictive performance. To select a small subset of variables, RFs are fitted iteratively. Specifically, at each iteration step a RF is grown and its importance (see Section 2.2) for classification is calculated. Variables that yield small VI scores are discarded subsequently. The elimination procedure is stopped when the number of remaining variables falls below a specific threshold or when the OOB accuracy is maximized (Diaz-Uriarte and Alvarez de Andres, 2006).

2.2 Importance

An essential standard feature of RF is that the importance of each predictor variable can be estimated. The RF approach serves two fundamentally different VI measures, the Gini importance and the permutation importance. The Gini importance of a predictor variable X_i is the total decrease in impurity $\Delta I = \Sigma_k \Delta i_k$. The Gini importance is obtained by adding up impurity decrease Δi_k of all nodes in a forest, where the corresponding predictor variable was selected for splitting. The Gini importance has been shown to be biased when the number of categories differs between predictor variables (Archer and Kimes, 2008; Strobl *et al.*, 2007). Moreover, bootstrapping observations without a replacement yields a less biased VI (Strobl *et al.*, 2007).

Another VI is the unscaled permutation importance, which is the mean decrease of accuracy for a predictor variable. This VI is calculated as follows: first, the prediction accuracy A_t is estimated for each tree t in forest T using OOB samples; second, the values of corresponding predictor variable are randomly permuted; third, prediction accuracy A_t^* is estimated using OOB samples again; and finally, the difference in accuracy, averaged over all trees in the forest, gives the unscaled permutation importance of the predictor variable

$$\overline{d} = \frac{1}{|T|}\sum_{t\in T} A_t - A_t^*. \tag{1}$$

The scaled permutation importance, often called z-score, is calculated by dividing the mean decrease of accuracy by its standard error over all trees in the RF

$$z = \frac{\overline{d}}{\sqrt{s^2/|T|}}. \tag{2}$$

The variance estimators differ between randomForest and RF in Fortran. As a result, both programs provide different scaled permutation importance scores. The estimator of randomForest is defined as

$$s^2 = \frac{1}{|T|}\sum_{t\in T} N_{\mathrm{OOB},t}(A_t - A_t^*)^2 - \overline{d}, \tag{3}$$

where $N_{OOB,t}$ determines the number of samples in OOB of the current tree. The variance estimator of RF in Fortran is defined as

$$s^2 = \frac{1}{|T|}\sum_{t\in T} (A_t - A_t^*)^2 - \overline{d}. \tag{4}$$

VI measures as described above can show a bias of correlated predictor variables such as SNPs in linkage disequilibrium (Meng *et al.*, 2009; Nicodemus and Malley, 2009; Nicodemus *et al.*, 2010; Strobl *et al.*, 2008). Permuting a predictor variable using the usual permutation scheme disrupts a potential dependency structure between the permuted variable and the other predictor variables. The disruption entails an inflation of the importance value of the predictor variable when predictors were associated with the outcome (Nicodemus *et al.*, 2010; Strobl *et al.*, 2008).

The conditional VI (CVI) is an approach to solve this problem (Strobl *et al.*, 2008). For preserving the dependency structure between a specific predictor variable and other predictor variables, the predictor variable in question is permuted only within groups of observations. Group assignment is determined by analyzing the corresponding dependency structure as described in detail by Strobl *et al.* (2008). It has been shown that the CVI reflects the importance of predictors of correlated predictors more reliably than usual importance measures (Strobl *et al.*, 2008). Therefore, the CVI should be applied to data that contain correlated predictor variables such as SNPs in linkage disequilibrium.

3 IMPLEMENTATION

3.1 Random jungle

The novel software package RJ implements all features of the reference implementation randomForest such as various tuning parameters, prediction of new datasets using previously grown forests, sample proximities and imputation. Commonly used VI measures are implemented, such as Gini importance, permutation importance and conditional importance measures. The features of RJ are shown in Supplementary Table 1. RJ additionally implements the variable backward elimination. When multiple CPU are available, RJ is able to perform RF on multiple CPUs simultaneously using multithreading and Message Passing Interface (MPI) parallelization.

RF in Fortran and randomForest grow ensembles of a CART, but the RF method is not restricted to a CART. Therefore, RJ serves a generalized framework for tree growing, which can be utilized to extend the set of tree types. RJ also implements CART, but several tree types such as conditional trees (Hothorn *et al.*, 2006) are currently under construction and will be added to RJ in the future.

RJ is written in the C++ language, and the program structure fundamentally differs from the randomForest and RF in Fortran implementations. A comparison of importance scores, computing time and memory consumption across different implementations is given in Section 3.2.

The software is freely available on www.randomjungle.org, where a detailed documentation of RJ can be found.

3.2 Comparison of importance values

A simulation study was set up for comparing importance score ranks of RJ with the reference implementation randomForest. To this end, we used the simulated data for rheumatoid arthritis (RA) that were provided for the Genetic Analysis Workshop (GAW) 15

(Miller *et al.*, 2007). Several loci contribute to the susceptibility, and nine major gene effects (Locus A–H and Locus DR) and three key covariate effects (smoking, age and sex) were simulated. The first replicate of the genome-wide SNP dataset, RA affection status and gender was utilized for the purpose of comparison. To mimic a case–control study, one affected sibling per affected pair for the cases and one unaffected sibling per control family for the controls were randomly selected. The dataset for application comprises 1500 cases and 2000 controls genotyped at 9187 SNPs. Ranks of Gini and unscaled permutation importance scores were investigated by applying RJ and randomForest to a subset of the GAW15 data. The subset comprised three informative predictor variables, i.e. sex, Locus C/DR (SNP6_153) and Locus D (SNP6_162), and three uninformative predictor variables (SNP2_394, SNP3_481, SNP1_98) that were randomly selected out of the set of all uninformative predictor variables (Miller *et al.*, 2007). Each program was applied 500 times in order to capture the variation of importance ranks. Data was analyzed using default parameter settings of 500 trees and $\mathrm{mtry} = \left\lceil \sqrt{M} \right\rceil = 3$. Finally, importance ranks of predictor variables yielded by randomForest and RJ were compared using boxplots. All applications were performed on computers running the SUSE Linux operating system with a 2.33 GHz Intel dual quad-core processor (8 CPUs) and 16 GB memory.

3.3 Performance and application to Crohn's disease

The real dataset was used to compare the different implementations and to find potential interactions.

The performance of RJ, randomForest, RF in Fortran and Willows was compared in terms of computing time and memory consumption of each software. RJ was run in two different modes, namely in a single CPU mode and in a 40-CPU mode using multithreading and MPI. All applications were performed on computers running the SUSE Linux operating system with a 2.33 GHz Intel dual quad-core processor (8 CPUs) and 16 GB memory. In the 40-CPU mode, five processes were distributed among five computers. Each process was performing on eight CPUs simultaneously using multithreading.

The real dataset is from a Crohn's disease GWA study, which has been described previously in detail (Duerr *et al.*, 2006). In brief, data of 513 Crohn's disease affected Caucasian cases and 515 Caucasian controls were analyzed. The samples were genotyped on the Illumina HumanHap300 Genotyping BeadChip (317 503 SNPs). The GWA study was funded by NIDDK IBD Genetics Consortium. Samples were visualized using MDS plots and outlying persons were excluded from MDS clusters by visual inspection of two experienced experts, resulting in 1006 persons (501 cases and 505 controls). Sex was the only covariate in the analysis in addition to the SNPs.

SNPs with a call rate <0.98 per study group, a MAF <0.05 in the cases and controls combined or a P-value <0.0001 for deviation from Hardy–Weinberg expectations in control group were excluded, resulting in 275 153 SNPs. The RJ software can handle missing data, i.e. imputing internally and analyzing data subsequently, but it is advised to impute data using standard imputing tools (Schwarz *et al.*, 2009). Missing genotypes were imputed by the IMPUTE program (Marchini *et al.*, 2007) using default parameters. Imputation uncertainty cannot be taken into account. Each implementation performed a RF analysis using the default forest size of 500 trees. To optimize the performance of the forest, the parameter mtry was tuned as recommended (Breiman and Cutler, 2004). The RF manual recommends choosing the mtry value that minimizes the OOB prediction error fraction. The parameter mtry was optimized by using several candidate values based on the formula $\mathrm{mtry} = \lfloor M/20 \cdot (1, \ldots, 19)' \rfloor$ is the number of predictor variables which are SNPs and sex. The results are shown in Supplementary Table 2. The minimal OOB prediction error was obtained for $\mathrm{mtry} = 247\,638$.

For investigating the genetic relevance of SNPs and their interactions, data analysis was performed by RJ using 100 000 trees in forest. The parameter mtry was optimized for 100 000 trees by comparing different values shown in Supplementary Table 2. The optimal mtry was found to be 27 515. The CVI was calculated for each SNP. For comparing results with a standard univariate method, the 275 153 SNPs were also analyzed using the common trend test, which tests SNPs for being associated with a disease (Ziegler and König, 2010). The P-values and their rank were compared with results of RJ analysis.

A network was created using the top 10 genes. The network was generated through the use of Ingenuity Pathways Analysis (Ingenuity Systems, www.ingenuity.com).

4 RESULTS

Comparison of importance scores shows that RJ and randomForest rank all variables in the same order. Results of Gini importance score and permutation importance scores investigation are shown in Supplementary Figure 1a, b and c. Both programs yield similar scores for all importance measures.

All implementations are able to handle the real dataset, but computing time and memory consumption differed substantially (Fig. 1). Specifically, the randomForest and RF in Fortran analyzed the dataset in 88.8 and 84.1 h, respectively, whereas RJ performed the same analysis in only 0.53 h using 40 CPUs in parallel. RF in Fortran is the fastest alternative tool. In comparison to RF in Fortran, RJ performed 159 times faster. With RJ running in a single CPU mode, a speed up of seven was still obtained. Willows required 1750 h for the analysis. RJ turned out to be the fastest program for real data analysis. The computing time of all implementations is depicted in Figure 1a.

The randomForest package and RF in Fortran consumed 9805 and 5421 MB memory, respectively (Fig. 1b). Considerably less memory was used by Willows, which consumed 136 MB memory. RJ spent 179 MB using one CPU. When using multiple CPUs, the program RJ distributed five processes among five computers using the MPI mode. Each process consumed 303 MB and utilized eight CPUs by using multithreading. RJ consumed more memory in the multi-processor mode because helping data structures have to be provided for every CPU.

The importance scores of the SNPs and their chromosomal positions are depicted in Figure 2. The two highest peaks are located on chromosomes 1 and 16, which correspond to genes *IL23R* and *NOD2,* respectively.

A comparison of positive CVI scores and two-sided P-values from the Cochrane–Armitage trend test is shown in Supplementary Figure 2. The smallest P-value of all positive CVI scores is 2×10^{-8}. The Pearson's correlation coefficient between scores and P-values is 0.38, showing a moderate association between importance scores

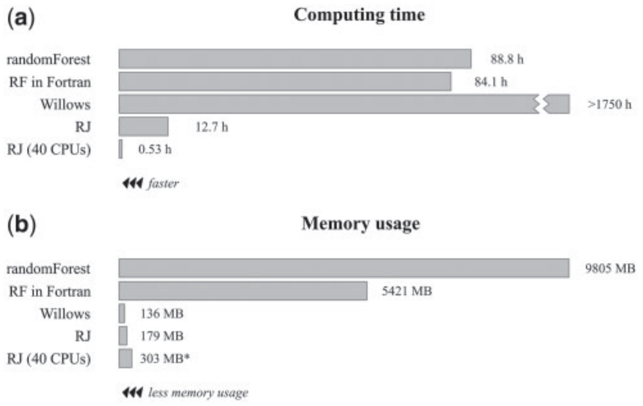

Fig. 1. Comparison of computing time and memory usage of several RF implementations. Each program analyzed a simulated dataset comprising 1006 samples genotyped at 275 153 SNPs. A short bar indicates a fast implementation of RF in comparison to other programs: (**a**) Comparison of computing time of five implementations. The figure reads for example: For analyzing data, RJ calculations took 0.53 h. (**b**) Comparison of memory usage of five programs. Memory was sparsely used by Willows and RJ in comparison to randomForest and RF in Fortran. (Asterisk indicates memory usage of each computer node.)

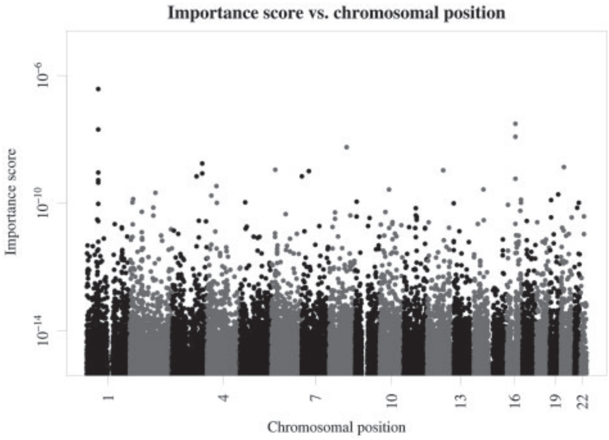

Fig. 2. CVI scores of SNPs and their chromosomal position. The axis of CVI scores was log transformed. Small and negative CVI values were omitted.

and *P*-values. Corresponding negative CVI scores are shown in Supplementary Figure 3.

The 10 most important genes for the Crohn's disease data are displayed in Table 1. The genes were derived by evaluating the most important SNPs that are located within genes as shown in Supplementary Table 3. The first 10 unique genes were selected for further investigations.

The four most important SNPs yielded the smallest *P*-values in the trend test. The interleukin 23 receptor (*IL23R*) and nucleotide-binding oligomerization domain containing 2 (*NOD2*) were found to be the most important genes by RJ analysis. *IL23R* and *NOD2* genes were also identified by several GWA studies and corresponding SNPs are known to be strongly associated with susceptibility to Crohn's disease (Barrett *et al.*, 2008; Duerr *et al.*, 2006; Rioux *et al.*, 2007; Wellcome Trust Case Control Consortium, 2007).

Table 1. Top 10 most important genes identified by RJ, which was performed on Crohn's disease GWA study data

Gene	rs-Number	CVI score	Rank of CVI score	*P*-value (trend test)	Rank of *P*-value
CEACAM4	rs5009916	5.98×10^{-5}	22	5.67×10^{-3}	2117
CDKAL1	rs9465994	1.30×10^{-4}	8	3.16×10^{-4}	158
CLSTN2	rs6439924	1.05×10^{-4}	14	2.58×10^{-5}	18
FBN3	rs4527136	5.01×10^{-5}	26	4.73×10^{-3}	1807
IL23R	rs11209026	1.63×10^{-3}	1	2.00×10^{-8}	1
	rs11465804	4.58×10^{-4}	3	1.44×10^{-7}	4
	rs1343151	1.19×10^{-4}	11	3.11×10^{-5}	29
	rs10889677	9.34×10^{-5}	16	8.22×10^{-6}	9
	rs2201841	8.65×10^{-5}	17	2.73×10^{-5}	19
NOD2	rs2066843	5.48×10^{-4}	2	1.44×10^{-7}	3
	rs2076756	3.65×10^{-4}	4	3.02×10^{-8}	2
	rs9465994	1.30×10^{-4}	8	3.16×10^{-4}	158
PRKG1	rs766208	6.97×10^{-5}	20	4.38×10^{-5}	37
PTPRD	rs1889820	4.76×10^{-5}	27	1.12×10^{-4}	71
SNX8	rs10950641	1.05×10^{-4}	13	3.33×10^{-5}	31
TNFSF10	rs9859259	1.12×10^{-4}	12	1.57×10^{-3}	661

The table displays the most important genes, rs-numbers of corresponding SNPs, CVI scores, ranks of CVI scores, trend test *P*-values and the rank of the *P*-values.

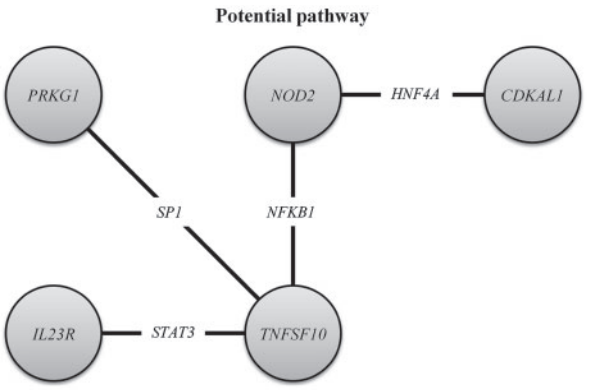

Fig. 3. Six of the 10 most important genes can be combined to a potential pathway. *TNFSF10* potentially interacts with *IL23R*, *NOD2* and *PRKG1*. The *NOD2* conceivably interacts with *CDKAL1*.

A moderate association of calsyntenin 2 (*CLSTN2*) ($P = 6 \times 10^{-5}$) with Crohn's disease was reported in literature (Rioux *et al.*, 2007). The tumor necrosis factor superfamily member 10 (*TNFSF10*) was high ranked by RJ analysis. The traditional trend test of *TNFSF10* yielded $P = 0.001576$. *TNFSF10* is involved in apoptosis and proliferation of human colon cancer cells (Baader *et al.*, 2005; Saaf *et al.*, 2007; Tang *et al.*, 2002; Tillman *et al.*, 2003). Colorectal cancer and Crohn's disease are related due to the fact that relative risk of colorectal cancers is significantly raised in Crohn's disease (Canavan *et al.*, 2006; Ekbom *et al.*, 1990).

Five of the 10 most important genes can be linked to a potential pathway by consulting additional genes, proteins and transcripts. The potential pathway is shown in Figure 3. The *IL23R* and

TNFSF10 potentially interact via the signal transducer and activator of transcription 3 (*STAT3*; Niu *et al*., 2001; Parham *et al*., 2002). *TNFSF10* possibly interacts with *PRKG1* via Sp1 transcription factor (*SP1*; Sellak *et al*., 2002; Xu *et al*., 2008). *TNFSF10* contingently interacts with *NOD2* via nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (*NFKB1*; Baetu *et al*., 2001; Gutierrez *et al*., 2002). Finally, *NOD2* conceivably interacts with CDK5 regulatory subunit associated protein (*CDKAL1*) via hepatocyte nuclear factor 4 alpha (*HNF4A*, Odom *et al*., 2004; Rioux *et al*., 2007).

5 DISCUSSION

The RF method was applied to genome-wide data using RJ and assigned a score of importance to each SNP. The resulting list of SNPs was investigated for potential interactions.

The results of the real data analysis validate the findings of GWA studies such as *NOD2* and *IL23R*. Results give also evidence of new potential interactions between genes that are associated with Crohn's disease. Specifically, the *TNFSF10* was not found to be strongly associated with Crohn's disease by traditional statistical tests. In contrast, RJ analysis detected that *TNFSF10* potentially interacts with *NOD2*, *PRGK1* and *IL23R*. *STAT3* is considered to be a link between *TNFSF10* and *IL23R*, which was shown to be moderately associated with Crohn's disease. The *TNFSF10* is involved in apoptosis of human colon cancer cells. *TNFSF10* might possibly explain a part of the high risk of colorectal cancers in Crohn's disease patients.

However, interpreting results and assessing biological plausibility is challenging. Results may show false positives. Thus, further investigation is needed to validate this specific causal relationship. Furthermore, although RF has the ability to detect very small main effects, its power to identify interactions depends on the presence of main effects. Thus, gene–gene interaction with no marginal effects might be left unrevealed when RF is applied.

Nevertheless, an impressive computational efficiency and memory management of RJ allow for analyzing high-dimensional data in an acceptable amount of time. Analyzing GWA data comprising thousands of observations and a million SNPs seem to be feasible with respect to time and memory consumption. The software computes up to 159 times faster than the fastest alternative implementation, and the program shows the same importance ranking with respect to the reference program. RJ presents various features such as RF growing, prediction, parameter tuning or imputation.

In summary, RJ is a promising software package for applying RF method to high-dimensional data such as GWA data. The application of RF to GWA data may help to identify potential interacting SNPs that were not found by traditional statistical approaches.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their helpful suggestions and comments. The NIDDK IBDGC Crohn's Disease GWA Study was conducted by the NIDDK IBDGC Crohn's Disease GWA Study Investigators and supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). This manuscript was not prepared in collaboration with Investigators of the NIDDK IBDGC Crohn's Disease GWA Study and does not necessarily reflect the opinions or views of the NIDDK IBDGC Crohn's Disease GWA Study or the NIDDK.

Funding: DFG (KO 2250/3-1); intramural funding from Medical Faculty of the University at Lübeck (E32-2009, SPP2); NIH grants 5RO1-HL049609-14, 1R01-AG021917-01A1; University of Minnesota; Minnesota Supercomputing Institute; GAW grant, R01-GM031575; ENGAGE (grant agreement number 201413); Atherogenomics (grant agreement number 01GS0831).

Conflict of Interest: none declared.

REFERENCES

Archer,K.J. and Kimes,R.V. (2008) Empirical characterization of random forest variable importance measures. *Comput. Stat. Data Anal.*, **52**, 2249–2260.

Baader,E. *et al.* (2005) Tumor necrosis factor-related apoptosis-inducing ligand-mediated proliferation of tumor cells with receptor-proximal apoptosis defects. *Cancer Res.*, **65**, 7888–7895.

Baetu,T.M. *et al.* (2001) Disruption of NF-kappaB signaling reveals a novel role for NF-kappaB in the regulation of TNF-related apoptosis-inducing ligand expression. *J. Immunol.*, **167**, 3164–3173.

Barrett,J.C. *et al.* (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.*, **40**, 955–962.

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. (2004) Random Forests 5.1. Available at http://www.stat.berkeley.edu/~breiman/RandomForests/cc_software.htm (last accessed date April 16, 2010).

Bureau,A. *et al.* (2005) Identifying SNPs predictive of phenotype using random forests. *Genet. Epidemiol.*, **28**, 171–182.

Canavan,C. *et al.* (2006) Meta-analysis: colorectal and small bowel cancer risk in patients with Crohn's disease. *Aliment Pharmacol. Ther.*, **23**, 1097–1104.

Chang,J.S. *et al.* (2008) Pathway analysis of single-nucleotide polymorphisms potentially associated with glioblastoma multiforme susceptibility using random forests. *Cancer Epidemiol. Biomarkers Prev.*, **17**, 1368–1373.

Cordell,H.J. (2009) Genome-wide association studies: detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **10**, 392–404.

Cox,T.F. and Cox,M.A.A. (2001) *Multidimensional Scaling*. *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton.

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

Duerr,R.H. *et al.* (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*, **314**, 1461–1463.

Ekbom,A. *et al.* (1990) Increased risk of large-bowel cancer in Crohn's disease with colonic involvement. *Lancet*, **336**, 357–359.

Gutierrez,O. *et al.* (2002) Induction of NOD2 in myelomonocytic and intestinal epithelial cells via nuclear factor-kappaB activation. *J. Biol. Chem.*, **277**, 41701–41705.

Hoh,J. *et al.* (2000) Selecting SNPs in two-stage analysis of disease association data: a model-free approach. *Ann. Hum. Genet.*, **64**, 413–417.

Hothorn,T. *et al.* (2006) Unbiased recursive partitioning. *J. Comput. Graph. Stat.*, **15**, 651–674.

Jakobsdottir,J. *et al.* (2009) Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. *PLoS Genet.*, **5**, e1000337.

Jiang,R. *et al.* (2009) A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics*, **10** (Suppl. 1), S65.

König,I.R. *et al.* (2008) Patient-centered yes/no prognosis using learning machines. *Int. J. Data Min. Bioinform.*, **2**, 289–341.

Liaw,A. and Wiener,M. (2002) Classification and Regression by randomForest. *R News*, **2**, 18–22.

Lunetta,K.L. *et al.* (2004) Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet.*, **5**, 32.

Macqueen,J.B. (1967) Some methods of classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathamatical Statistics and Probability*, University of California Press, Berkeley and Los Angeles, California, pp. 281–297.

Manolio,T.A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

Marchini,J. *et al.* (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.*, **37**, 413–417.

Marchini,J. *et al.* (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, **39**, 906–913.

McCarthy,M.I. *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356–369.

McKinney,B.A. *et al.* (2006) Machine learning for detecting gene-gene interactions: a review. *Appl. Bioinformatics*, **5**, 77–88.

McKinney,B.A. *et al.* (2009) Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLoS Genet.*, **5**, e1000432.

Meng,Y. *et al.* (2009) Performance of random forest when SNPs are in linkage disequilibrium. *BMC Bioinformatics*, **10**, 78.

Miller,M.B. *et al.* (2007) Genetic Analysis Workshop 15: simulation of a complex genetic model for rheumatoid arthritis in nuclear families including a dense SNP map with linkage disequilibrium between marker loci and trait loci. *BMC Proc.*, **1** (Suppl. 1), S4.

Moore,J.H. *et al.* (2010) Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, **26**, 445–455.

Nicodemus,K.K. and Malley,J.D. (2009) Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics*, **25**, 1884–1890.

Nicodemus,K.K. *et al.* (2010) The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, **11**, 110.

Niu,G. *et al.* (2001) Overexpression of a dominant-negative signal transducer and activator of transcription 3 variant in tumor cells leads to production of soluble factors that induce apoptosis and cell cycle arrest. *Cancer Res.*, **61**, 3276–3280.

Odom,D.T. *et al.* (2004) Control of pancreas and liver gene expression by HNF transcription factors. *Science*, **303**, 1378–1381.

Parham,C. *et al.* (2002) A receptor for the heterodimeric cytokine IL-23 is composed of IL-12Rbeta1 and a novel cytokine receptor subunit, IL-23R. *J. Immunol.*, **168**, 5699–5708.

Province,M.A. *et al.* (2001) Classification methods for confronting heterogeneity. *Adv. Genet.*, **42**, 273–286.

R Development Core Team (2009) R: a language and environment for statistical computing. Available at http://www.r-project.org (last accessed date April 16, 2010).

Rioux,J.D. *et al.* (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.*, **39**, 596–604.

Saaf,A.M. *et al.* (2007) Parallels between global transcriptional programs of polarizing Caco-2 intestinal epithelial cells in vitro and gene expression programs in normal colon and colon cancer. *Mol. Biol. Cell.*, **18**, 4245–4260.

Samani,N.J. *et al.* (2007) Genomewide association analysis of coronary artery disease. *N. Engl J. Med.*, **357**, 443–453.

Schapire,R.E. (1990) The strength of weak learnability. *Mach. Learn.*, **5**, 197–227.

Schwarz,D.F. *et al.* (2007) Picking single-nucleotide polymorphisms in forests. *BMC Proc.*, **1** (Suppl. 1), S59.

Schwarz,D.F. *et al.* (2009) Evaluation of single-nucleotide polymorphism imputation using random forests. *BMC Proc.*, **3**, S65.

Sellak,H. *et al.* (2002) Sp1 transcription factor as a molecular target for nitric oxide- and cyclic nucleotide-mediated suppression of cGMP-dependent protein kinase-Ialpha expression in vascular smooth muscle cells. *Circ. Res.*, **90**, 405–412.

Strobl,C. *et al.* (2008) Conditional variable importance for random forests. *BMC Bioinformatics*, **9**, 307.

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

Sun,Y.V. *et al.* (2007) Classification of rheumatoid arthritis status with candidate gene and genome-wide single-nucleotide polymorphisms using random forests. *BMC Proc.*, **1** (Suppl. 1), S62.

Tang,X. *et al.* (2002) Cyclooxygenase-2 overexpression inhibits death receptor 5 expression and confers resistance to tumor necrosis factor-related apoptosis-inducing ligand-induced apoptosis in human colon cancer cells. *Cancer Res.*, **62**, 4903–4908.

Tillman,D.M. *et al.* (2003) Rottlerin sensitizes colon carcinoma cells to tumor necrosis factor-related apoptosis-inducing ligand-induced apoptosis via uncoupling of the mitochondria independent of protein kinase C. *Cancer Res.*, **63**, 5118–5125.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Xu,J. *et al.* (2008) Sp1-mediated TRAIL induction in chemosensitization. *Cancer Res.*, **68**, 6718–6726.

Zhang,H. *et al.* (2009) Willows: a memory efficient tree and forest construction package. *BMC Bioinformatics*, **10**, 130.

Ziegler,A. *et al.* (2007) Data mining, neural nets, trees–problems 2 and 3 of Genetic Analysis Workshop 15. *Genet. Epidemiol.*, **31** (Suppl. 1), S51–S60.

Ziegler,A. and König,I.R. (2010) *A Statistical Approach to Genetic Epidemiology: Concepts and Applications*. Wiley-VCH, Weinheim.