BRIEFINGS IN BIOINFORMATICS. VOL 14. NO 2. 251–260
Advance Access published on 18 May 2012
doi:10.1093/bib/bbs024

Machine learning approaches for the discovery of gene–gene interactions in disease data

Rosanna Upstill-Goddard, Diana Eccles, Joerg Fliege and Andrew Collins

Submitted: 14th March 2012; Received (in revised form): 20th April 2012

Abstract

Because of the complexity of gene–phenotype relationships machine learning approaches have considerable appeal as a strategy for modelling interactions. A number of such methods have been developed and applied in recent years with some modest success. Progress is hampered by the challenges presented by the complexity of the disease genetic data, including phenotypic and genetic heterogeneity, polygenic forms of inheritance and variable penetrance, combined with the analytical and computational issues arising from the enormous number of potential interactions. We review here recent and current approaches focusing, wherever possible, on applications to real data (particularly in the context of genome-wide association studies) and looking ahead to the further challenges posed by next generation sequencing data.

Keywords: *machine learning; gene–gene interaction; random forest; support vector machines; multifactor-dimensionality reduction; genome-wide association study*

INTRODUCTION

Genes influence all human diseases and yet much of the genetic landscape of many common diseases is still uncharacterized. Genome-wide association studies (GWAS) using single nucleotide polymorphisms (SNPs) have been extensively used to uncover genetic architecture [1] by testing variants individually for association with particular diseases or traits [2, 3]. However, GWAS have explained only a small proportion of the genetic variation underlying disease [1, 4]. For common diseases the effect of an individual SNP on disease susceptibility is generally small and emerging evidence suggests that many low-penetrance variants interact multiplicatively [5] with increasing numbers of risk alleles contributing to significantly elevated disease risks [6]. Therefore, it is likely that much of the genetic variation underlying common diseases arises through interactions between many genes and environmental factors; a form of epistasis [7]. Thus the identification of individual disease-related SNPs may be less useful for disease prediction than the identification of the epistatic relationships underlying genetic disease.

The term epistasis has been used to refer to at least two phenomena which may be related in complex ways. Biological epistasis, which occurs at the cellular level, corresponds to the physical interactions amongst biomolecules in gene regulatory networks and pathways that impact on phenotype. Hence, the impact of a gene on an individual's phenotype depends on one or more additional genes. Alternatively, statistical epistasis reflects differences

Corresponding author. Andrew Collins. Genetic Epidemiology and Bioinformatics, Faculty of Medicine, University of Southampton, Duthie Building (808), Southampton General Hospital, Tremona Road, Southampton SO16 6YD. Tel: 44 (0) 2380796939; Fax: 44 (0) 2380794264. E-mail: arc@soton.ac.uk

Rosanna Upstill-Goddard is studying for a Breast Cancer Campaign funded PhD at the University of Southampton focusing on machine learning approaches applied to early-onset breast cancer data.

Diana Eccles is Professor of Cancer Genetics at the University of Southampton and is undertaking research into early onset breast cancer using the 'Prospective study of Outcomes in Sporadic versus Hereditary breast cancer' (POSH) cohort.

Joerg Fliege is head of the Operational Research Group in the School of Mathematics at the University of Southampton and is involved in developing new mathematical tools and techniques in optimization and machine learning.

Andrew Collins is head of the Genetic Epidemiology and Bioinformatics Research Group at the University of Southampton and is involved in association and sequencing studies for a number of diseases for which machine learning methods show promise.

in biological epistasis among a population of individuals: the deviation from additivity within a statistical model of the relationship between multiple genotypes and phenotype(s) at a population level [1, 8]. Moore and Williams [8] present conceptual relationships between biological and statistical epistasis diagrammatically in their Figure 2. Phillips [9] has suggested that epistasis can be split into three categories: compositional epistasis, functional epistasis and statistical epistasis. Compositional epistasis is introduced to represent the traditional definition of epistasis as the blocking of the effect of an allele by an allele at another locus. However defined, the relationships between biological and statistical forms of epistasis are complex and statistical interaction does not necessarily reflect interaction on a biological level [10].

One of the major problems associated with uncovering epistatic interactions is the volume of data to be analysed; as the number of SNPs increases the number of potential interactions increases exponentially [7], known as the 'curse of dimensionality'. The potential complexity of such interactions supports the use of machine learning and data mining techniques. Machine learning (ML) approaches employ algorithms to 'learn' from training data sets to solve problems and enable predictions about outcomes in other data based on patterns and rules learned. There are several issues that need to be considered when developing ML methods for the identification of epistasis including: genetic heterogeneity (which may be common in complex diseases[11]), the presence (or absence) of main effects, and the number of SNPs involved in the interactions (which is usually unknown in advance) [11].

EARLY ML APPROACHES

A range of ML methods have been developed over the past 10–15 years with the aim of uncovering gene–gene interactions implicated in common complex diseases. Here we discuss some approaches that have been used to detect epistasis, namely multifactor-dimensionality reduction (MDR), neural networks (NNs), random forest (RF), and support vector machines (SVMs).

Multifactor-dimensionality reduction

MDR was one of the first ML methods developed to detect and characterize gene–gene interactions [12, 13]. In the first stage of MDR, *n* genetic factors (e.g. SNPs) are selected from the entire set of factors. All possible multifactor (SNP genotype) combinations are represented in cells in *n*-dimensional space and each cell is assigned a case-control ratio. Multilocus genotypic predictors are thus reduced from *n* dimensions to one dimension by classifying each cell as either low-risk or high-risk, based on a threshold value of cases-to-controls [12, 14]. Following classification cross-validation is carried out to estimate the prediction error of each model by splitting the data into a training set consisting of 90% of the data and a testing set of the remaining 10%. A model is developed based on the classification of genotypes in the training set which is used to predict disease status of genotypes in the test set. The cross-validation process is repeated 10 times and the prediction error is averaged [12]. MDR modelling can thus be applied to real disease data to search for epistasis and any predictors designated as 'high-risk' are, therefore, potentially disease-related. This approach was evaluated using a sporadic breast cancer data set [12]. A statistically significant high-order interaction was detected amongst four polymorphisms in the absence of any significant main effects, one of the earliest reports of such an interaction associated with a common multifactorial disease. The power of MDR was found to be robust to the presence of 5% genotyping error, 5% missing data and a combination of the two for a number of different two-locus epistasis models. Additional advantages of using MDR for the discovery of epistasis include:

(i) The model-free approach, invaluable for diseases such as sporadic breast cancer for which the mode of inheritance is unknown and likely to be complex.
(ii) The capability of MDR for detecting and characterizing multiple genetic loci simultaneously and, through the use of cross-validation, minimizing the false-positive rate.
(iii) The number of interaction terms does not grow exponentially as each new variable is added [12].

However, some disadvantages associated with this method impact upon its reliability as a predictor of disease–genotype interactions. In the presence of a high (50%) phenocopy–genetic heterogeneity rate, power is greatly compromised [14] supporting the need for refinements to effectively deal with genetic heterogeneity in complex trait data. The resulting

models can be difficult to interpret [12]—although genotypes are classified as 'high-risk' or 'low-risk' there is no quantitative assessment of how high- or low-risk they are, thus it is difficult to determine which of the putative interactions are most likely to be disease-related and warrant further investigation. MDR (and extensions to MDR) have only been successful when applied to a small number of SNPs in certain genes of (known) interest [12, 13, 15, 16]. The MDR approach alone is not directly applicable to GWAS data, given the huge number of interactions to be assessed; however, using a filter algorithm to isolate a subset of potentially interesting SNPs for MDR analysis can overcome this limitation. Finally, MDR has a high false positive and negative error rate when the case and control ratio in a genotype combination is closely similar to that in the whole data set[15].

Neural networks

NNs were originally developed to model neurons but are now regularly used for data mining in a wide range of fields [17, 18] with 'feed-forward/back-propagation' networks being the most common [19]. They have excellent power for performing pattern recognition and classification [11] and are capable of dealing with voluminous data [18]. A NN resembling a directed graph where the nodes represent genetic elements (SNPs), and the arcs are the connections (interactions) between the elements, has been developed for genetic applications [18]. The nodes are arranged into layers. One or more nodes reside in the input layer and receive the information to be processed by the NN. The input layer links to multiple nodes in a hidden layer (of which there may be several) via arcs. Finally, there is an output node. Each arc is assigned a weight which, initially, are chosen randomly, but through training the network on test data, weights are adjusted to minimize the error rate [19]. The target of the NN is the recognition of corresponding patterns in real data, based on patterns observed in test data, and for predictions about patterns not seen before through recognizing sub-patterns and correlations in the data [19]. To uncover genetic loci potentially involved in epistatic interactions NNs are trained using known genotypes as inputs and known phenotypes as outputs and the development of the internal weighting structure is of particular importance. The internal weight structure of the network can be analysed after training to determine the effect of each locus on the resulting phenotype [19].

NN applications to disease data have shown variable success. Motsinger-Reif *et al.* [18] suggest that this may be due to the use of sub-optimal NN architecture. Exhaustive search of all possible architectures to find the optimal structure is infeasible and so one solution is to optimize architecture with ML algorithms. Examples of such algorithms include the Genetic Programming optimized NN (GPNN) [20, 21] and Grammatical Evolution NN (GENN) [18] [using genetic programming (GP) or grammatical evolution (GE) respectively to optimize a NN]. GP aims to 'evolve' computer programs to solve complex problems [22]. First, an initial population of randomly generated computer programs is produced. Each program is run on a problem and assigned a fitness value based on its performance. The best programs are chosen to go forward for 'reproduction' following the 'survival of the fittest' principle. Some programs are taken into the next generation unaltered, while others undergo 'crossover' in which new programs are created from combinations of components of the original programs. This procedure is repeated for a number of generations to find the optimal program [23]. GE is a variation of, and improvement on, GP, with more flexibility[17]. GE uses populations consisting of linear genomes which constitute individuals. Each genome is divided into codons which are translated into phenotypes (the NN) by the grammar [17]. In a similar way to GP, the resulting phenotypes can be tested for fitness and subsequent generations produced to find the optimal model. GPNN has higher power to detect gene–gene interactions in the presence of non-functional SNPs than the more traditional Back Propagation NN (BPNN) [23] while power comparisons have shown that GENN consistently outperforms GPNN [17, 18]. NNs can screen out loci that do not affect the phenotype, thus reducing the number of genetic locus combinations to be tested [19]. Network approaches can also be used to identify genetic interactions through exhaustive enumeration of all possible pairwise interactions; however, this approach only searches for SNPs with strong pairwise interactions so may overlook SNPs with higher order interactions [24].

Genetic heterogeneity, polygenic inheritance, high phenocopy rates, and incomplete penetrance are problematic in the search for epistasis. Some of the characteristics of NN methods render them capable of addressing these difficulties; pattern

recognition is well suited to address genetic heterogeneity and polygenic inheritance while signal filtering addresses high phenocopy rates and incomplete penetrance [19].

Random forest

RF are a type of high-dimensional non-parametric predictive model composed of a collection of classification or regression trees [25] generated from random vectors [26]. Each tree of a RF is grown from a training set (or bootstrap sample) from the original data using random feature selection and trees are grown to their full extent without pruning. The bootstrap sample of size n is produced from the original sample, also size n, with variables chosen with replacement. Thus some variables will be chosen multiple times while others will not be chosen at all [25].The best split at each node in each tree is chosen from a random subset of the predictor variables [27]. The so-called 'out-of-bag' (OOB) estimates of prediction error are then generated from the observations that are not chosen in the bootstrap sample (often up to one third of cases are not included). The RF algorithm is an effective prediction tool with the potential to uncover interactions among genes that do not exhibit strong main effects [22], however, it has been suggested that their ability to detect interactions actually depends on the presence of main effects, no matter how weak [28]. Thus, this approach may lack power to uncover those interactions that occur in the absence of any main effects.

A recent study used the RF approach to uncover interacting SNPs contributing to rheumatoid arthritis, but no significant interactions were found that could be replicated in a follow-up cohort [29]. Power calculations have further indicated that this method will only detect those interactions with a large effect size [29]. However, an advantage of RFs is that they do not 'overfit' the data, and, as the number of trees in the RF increases, the prediction error converges to a limiting value [26]. An importance score is provided for each variable in a RF [27] rendering it capable of identifying SNPs predictive of a phenotype. This has prompted suggestions that RFs could be used to highlight significant SNPs for analysis with other methods [25]. However, this would conflict with the suggestion that RFs are useful tools to uncover genetic epistasis since the detection of interactions between variables is more important than the effect of single SNPs on disease status. A further downside of the RF method is that, although it has shown considerable promise in low-dimensional data (~100 SNPs and 10,000 observations), it has not been successfully applied to GWAS data [28].

Support vector machines

SVMs are classification techniques which are potentially as powerful as NNs [30]. In the development of a supervised learning approach the actual outcome of the (training) data is given and similar patterns are searched for during testing [31]. In its simplest form, a SVM is focused on identifying a linear separator to divide data points of two classes and is thus a non-probabilistic binary linear classifier. Furthermore, using kernel functions, non-linear separators can be established by modifying the input space. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. SVMs have shown excellent power to detect epistasis in both simulated and real datasets [11, 31]; Listgarten *et al.* [31] identified variants in a number of genes associated with breast cancer risk. A quadratic kernel was used and the authors showed that multiple SNP sites from several genes at distant parts of the genome were better at identifying breast cancer patients than single SNPs. When compared to MDR this approach provides more interpretable output; however, unlike MDR, SVMs cannot cope well with missing data [11]. Chen *et al.* [11] employed an SVM approach which was combined with search algorithms to produce four different models to detect epistasis, in the absence of genetic heterogeneity. Sparse SVMs [32] have been developed to select variables for inclusion in the model as a pre-processing step. This technique aims to reduce the instabilities in SVM results that arise from small changes in training/validation data. Such an approach might be usefully applied to the study of epistasis.

MORE RECENT MODELS AND APPLICATIONS

Recently a number of new ML methods have been proposed which utilize different approaches to detect epistasis [13, 15, 16, 27, 28, 33–35], some based on the methods already discussed while others introduce new approaches.

Extensions to MDR

MDR has proven a popular method for identifying epistasis. Recently, new MDR models have been developed to address some of the limitations of the original methods, some of which have been successfully applied to a range of genetic association studies [13, 15, 16, 34, 35]. It was previously demonstrated that MDR cannot effectively model epistasis when genotype combinations have case-control ratios similar to those in the full data set. However, robust MDR (RMDR) was proposed to deal with this limitation. In traditional MDR if the case-control ratio of a cell equals that of the whole data set, or a cell is empty, then it is randomly assigned high- or low-risk status. In robust MDR genotypes are pooled into three groups: high-risk, low-risk, or unknown-risk—based on the statistical significance of association of each multilocus cell with case-control status. If a cell has case-control ratio equal (or close) to that of the whole data set, then it is labelled unknown-risk and is excluded from the model. Results are thus simplified and easier to interpret than traditional MDR. The RMDR approach was evaluated with a bladder cancer dataset and confirmed findings of the MDR approach but with a simpler, easier-to-interpret model which was more computationally efficient [13].

Quantitative classification of SNP–SNP interactions is absent from traditional MDR, making it difficult for researchers to establish which interactions are potentially the most important. Two recently proposed methods designed to produce more definitive results are OR-MDR and MB-MDR. Chung *et al.* [15] proposed Odds Ratio-based MDR (OR-MDR) which uses the same method as MDR to categorize genotypes as high-risk or low-risk but includes the odds ratio for each genotype combination. A high odds ratio indicates an interaction that is potentially more high-risk compared to an interaction with a low odds ratio, thus the most high-risk interactions can be easily identified with such an approach. Model-based MDR (MB-MDR) [34, 35] assigns genotypes into three categories: high-risk, low-risk, or no evidence and all three are tested for association with the disease/trait. This test is designed to reduce the risk of missing of important interactions and the inability to deal effectively with main effects and confounding factors. A two-stage approach is utilized to uncover interactions: a synergy measure among potentially interacting genes is used in the first stage and MB-MDR is used in the second stage. This method aims to produce more accurate identification of which interactions are high-risk and results show that, when compared to MDR, it identifies far fewer interactions as high- or low-risk. Many interactions that are classified as high-risk by MDR are assigned to the 'no evidence' group in MB-MDR. Evidence suggests that this is a smaller set of more reliable results and some of the genotypes assigned to the 'no evidence' group are in fact disease-related and should not be excluded from the model. The inclusion of this group seems to be particularly important in scenarios where minor allele frequencies (MAFs) are low, where genetic heterogeneity is present and when there is less power to make reliable statements about the risk status of the genotypes.

To address MDR's inability to take into account covariates such as age, sex and ethnicity, and inability to deal with continuous data, Generalized MDR (GMDR) has been proposed. GMDR is similar to MDR in that it splits genotypes into the two classifications of risk; however, it uses a score based on the maximum-likelihood estimates for each variable, rather than the case-control ratio. The score is calculated with the inclusion of terms for covariates and dichotomous/continuous data. GMDR significantly increases the accuracy of risk prediction when the data contains covariates and is applicable to balanced case-control or random data [16].

Extensions to RF

SNPInterForest [27] is based on the RF approach but is more successful at uncovering disease-associated SNPs and has the capacity to simultaneously identify multiple interactions. SNPInter Forest is more sensitive to SNPs with limited marginal effect, something the original RF algorithm performs poorly with. A modification to the RF method is introduced to prevent the importance scores of SNPs without marginal effects from being underestimated. Multiple-SNP selection occurs at each node, in contrast to the original RF algorithm in which only a single SNP is used. This significantly improves the ability of this approach to detect SNPs associated with disease. It can be challenging to extract useful biological information from RF analyses with respect to biological interactions. SNPInter Forest addresses this issue by evaluating the interaction strength of SNP combinations. Each branch of a tree represents a possible SNP interaction on that branch and, if a certain SNP combination appears

more often on a branch, then those SNPs are likely to interact more strongly. Interaction strength is calculated from the number of times each SNP combination appears in each branch of each tree. Normalisation is applied in order to identify the weaker interactions and those interactions due to single SNPs with strong associations. SNPInterForest also demonstrates high recall rates and low false positive rates, however, it is very computationally demanding [27].

SNPInterForest has been identified as outperforming other methods such as 'BOolean Operation-based Screening and Testing' (BOOST). BOOST is a computationally efficient two-stage statistical method applied to analyse all pairwise interactions in genome-wide data [36] and in simulated data with weak marginal effects. In the absence of marginal effects, BOOST has been shown to produce many more false positive results. The ability of SNPInterForest to detect high order epistatic interactions between more than two SNPs was also assessed in simulated data. Five datasets were produced using a model of three SNPs, two of which are moderately associated with the disease by a pure epistatic interaction and a third SNP with a weaker effect that amplifies the interactive effect. SNPInterForest successfully identified the interactions in all five datasets.

RFs are often used for the selection of a subset of variables [22] rendering them useful for identifying potentially interesting SNPs in a two-stage approach [33]. For example, the TRM method [33] uses RF to identify and select important variants and Multivariate Adaptive Regression Splines (MARS), a nonparametric regression method, to detect interactions. RFcouple [37] on the other hand has been suggested as a pre-screening method for MDR. The advantage of two-stage approaches is that a subset of potentially significant SNPs is selected by a filter algorithm and a ML approach is employed to search for potential interactions; thus a smaller set of potentially interesting SNPs can be exhaustively searched for epistasis [3, 38].

Lin *et al.* [33] combined both RF and MARS in the TRM approach because neither method alone was considered optimal for selecting the optimal genotype combination for predicting phenotypes in studies with large numbers of SNPs. Individually RFs can have difficulties revealing underlying interaction patterns and MARS can have difficulty coping with numerous non-functional SNPs. In this study three approaches were compared: TRM_{OOB}, TRM_{IS}, and MARS alone. TRM_{OOB} is a version of TRM that uses RF_{OOB} and MARS while TRM_{IS} uses RF_{IS} and MARS. RF_{OOB} is based on the unused 'OOB' observations and RF_{IS} is based on the importance spectrum of the original data compared with that of the permuted data. TRM_{OOB} demonstrated higher true positive rates and lower false positive rates than the other two approaches in a simulation study with 100 SNPs. TRM has not yet been applied to a large dataset, such as a GWAS, nor has it been compared to other more established methods with proven success in the past. The study has, however, demonstrated a two-stage approach for screening and testing SNPs as capable of uncovering potential interactions.

De Lobel *et al.* [37] propose using RFcouple as a pre-screening method for MDR. RFcouple is based on RFs but uses information from the ratio of cases to controls for each genotype to define a new variable for each SNP pair. Thus the data set contains a variable for each SNP pair rather than for each individual SNP. An RF is constructed based on these data and SNP pairs are chosen based on Z-scores, which are related to prediction error and standard error of the RF in the permuted data. The single SNPs that make up these pairs are retained and then analysed by a method such as MDR. Power when RFcouple is used before MDR is always comparable to or greater than using MDR alone.

Random Jungle (RJ) is an implementation of the RF method which is aimed at analysing data on a genome-wide scale, i.e. 1000s of SNPs [28]. Application of RJ to Crohn's disease GWAS data confirmed previous GWAS findings as well as uncovering new interactions between Crohn's-associated genes. RJ is much more computationally efficient than other RF implementations allowing for feasible analysis of high-dimensional GWAS data in a realistic time frame. However, RJ differs from many methods in that it tests association allowing for interaction, rather than testing directly for interaction [36]. In line with the traditional RF approach, RJ has difficulty in detecting interactions when SNPs only have weak main effects; the trees are constructed based on the main effects of SNPs [36], thus such an approach is not useful in the absence of main effects. Table 1 provides an overview of some of the main ML approaches used to detect gene–gene interactions and some of their strengths and limitations.

Table 1: Strengths and limitations of some machine learning approaches

Method	Strengths	Limitations
MDR	Detects multiple genetic loci simultaneously, keeping false-positive rate low. Model-free—important when mode of inheritance is unknown. Nonparametric—the number of interaction terms does not grow exponentially as each new variable is added. Power remains high with 5% genotyping error and/or 5% missing data for various two-locus epistasis models. Cross-validation minimizes false positive rate.	Power significantly reduced with high (50%) phenocopy/genetic heterogeneity. No quantitative assessment of each model to determine which is the most high-risk—models difficult to interpret. Can be computationally intensive, particularly when the number of SNPs to be evaluated exceeds 10. False positive/negative error rates high when case-control ratio in test data is close to that in the whole dataset. May identify totally different models influenced by missing values in the data.
RMDR	Produces easier to interpret models than MDR—classifies genotypes with a case-control ratio close to the whole data set as 'unknown risk' (excluded from model).	Significant computational burden. Takes longer to evaluate one-way, two-way and three-way models than MDR.
OR-MDR	Like MDR but with odds ratio for each genotype combination—a quantitative measure of disease risk. Provides confidence interval for each genotype combination.	Cannot classify an empty cell. Computationally expensive, particularly when the number of SNPs exceeds 10. Impact of genetic heterogeneity/phenocopy unknown.
MB-MDR	Genotypes classified as low-risk/high-risk/no evidence—reducing number of interactions classified as high-risk. Genotypes in the 'no evidence' group potentially disease related and considered in the model. Improved power and false positive rate compared to MDR.	Phenocopy and genetic heterogeneity significantly reduce power.
GMDR	Uses score based on maximum-likelihood (ML) rather than case-control ratio. ML score includes covariates-significantly increasing accuracy of risk prediction.	Like MDR genotypes only assigned to two risk groups with no quantitative assessment. Can be computationally intensive, as above.
Neural Networks (NNs)	Excellent power for pattern recognition/classification Capable of dealing with large volumes of data. Accommodates genetic heterogeneity/polygenic inheritance/high phenocopy rates/incomplete penetrance.	It is impossible to enumerate all possible NN architectures and altering the architecture can change results of data analyses. Thus there is no way to be certain that the architecture being used is optimal.
GPNN	GP optimizes the architecture of NN. High power to detect interactions in the presence of non-functional SNPs. Preferable when functional SNPs are unknown and variable selection as well as model fitting required. Does not 'overfit' data. High power in epistasis model with weak marginal effect. Modelling flexibility—no need to select optimal inputs, weights, connections, or hidden layers.	High false positive rates in three locus models. Requires a parallel processing environment. The output is a binary expression tree which can be large (up to 500 nodes) and difficult to interpret.
GENN	GE optimizes the NN architecture. Consistently outperforms GPNN—optimizes NN more efficiently in fewer generations than GP. High power to detect risk loci in complex disease.	
RF	May uncover interactions among genes that do not exhibit strong main effects. Does not 'overfit' the data and prediction error converges to a limiting value. Identifies SNPs predictive of a phenotype.	Ability to detect interactions depends on main effects, no matter how weak. No demonstrated success in GWAS data. Sometimes underestimates importance scores of SNPs without marginal effects. Can be challenging to extract useful biological information. Only detects interactions with large effect size.
SNPInterForest	Simultaneously identifies multiple interactions. Does not underestimate importance scores of SNPs without marginal effects. Multiple SNP selection at each node improves ability to detect disease SNPs even when marginal effects absent. Evaluates interaction strength of SNP combinations. Demonstrates high recall/low false-positive rates. Interactions found in presence of genetic heterogeneity.	Very computationally intensive.

(continued)

Table 1 Continued

Method	Strengths	Limitations
TRM	A subset of potentially interesting SNPs can be searched exhaustively for interactions.	Has only been applied to small data sets (100 SNPs). Has not been compared to more established methods.
RJ	Designed to analyse data on a genome-wide scale. More computationally efficient than RF implementations—analysis of high-dimensional GWAS data feasible.	Tests association allowing for interaction rather than testing directly for interactions. May fail to detect interactions when only weak main effects.
SVM	More interpretable output compared to MDR. Readily generalized to new data structures. No user-defined decisions required for classification.	May not cope well with missing data. Power reduced in the presence of genetic heterogeneity.

LIMITATIONS OF CURRENT MODELS AND FUTURE DIRECTIONS

There are many difficulties associated with the detection of epistasis in GWAS related to both the data to be analysed and the capabilities of the ML methods being used. Firstly, there is the complexity of the disease data which includes allelic/locus heterogeneity, phenocopies, trait heterogeneity, phenotypic variability [38] and incomplete penetrance [10]. Some of the models discussed here have been developed to deal with such limitations. For example, it has been suggested that RF methods may be successful at dealing with certain types of heterogeneity [22, 27, 28], while some of the characteristics of NNs render them capable of addressing genetic heterogeneity, polygenic inheritance, high phenocopy rates and incomplete penetrance [17, 19].

Secondly, the computational burden associated with the search for gene–gene (SNP–SNP) interactions is potentially huge [39], particularly when searching for interactions between two or more SNPs within a GWAS [3]. The majority of methods discussed have demonstrated success in simulated data containing, at most, a few hundred SNPs. While such results are promising, it is currently unclear how successful some of these methods will be at dealing with up to 500 000 SNPs in a GWAS. Aside from the computational burden, the outputs may present serious challenges for biological interpretation. The power of many methods is significantly reduced when attempting to uncover higher order interactions. The issue of designing 'sufficiently powerful' studies has received relatively little attention to date. Although an exhaustive search of pairwise interactions in GWAS data might become computationally feasible extensive validation of candidate interactions in independent samples is essential, as in GWAS generally, to confirm or refute discoveries. More sophisticated approaches capable of modelling higher order interactions need to be developed but may require the use of expert knowledge of biological and biochemical pathways to choose SNPs likely to be associated with a particular disease [22]. It may also be necessary, and more powerful, to employ a two-stage model, in which filter algorithms select a subset of SNPs and a ML method exhaustively searches for interactions [3, 38, 40]. This approach may be less time-consuming and produce easier-to-interpret models [40]. However, some argue it is likely that SNPs with strong epistasis but weak main effects will be filtered out [36], so these methods will not necessarily find the optimal solution. Moreover, it is often the case that individual SNPs are assessed for disease association based on an importance score that does not take into account interactions with other SNPs. Clearly, when searching for epistasis, it is the interactions between SNPs that are of importance. Thus, a SNP with a high importance score but no involvement in SNP–SNP interactions is clearly not useful in this context.

There may be little similarity between biological and statistical epistasis; biological epistasis occurs at the cellular level within an individual while statistical epistasis addresses genetic variation on a population scale [1]. Most methods, however, test statistical, rather than biological, epistasis [36].

Finally, all approaches discussed are successful at uncovering epistasis in simulated data, with some also being successful in application to disease data of varying volume and complexity (Table 2). However, most studies have focussed on validating previously produced results with few actually uncovering new disease related interactions. While it is important that methods are tested on real data

Table 2: Some applications of machine learning approaches to genetic data

Method	Successful scenario	Reference
MDR	Applied to 10 polymorphisms in five genes related to oestrogen metabolism in breast tissue. A four-locus interaction associated with risk for sporadic breast cancer identified.	[12]
RMDR	Study examined the relationship between DNA repair gene SNPs, smoking and bladder cancer. Seven SNPs in five genes involved in DNA repair tested. Verified results from an MDR study using the same data but provided a much clearer model of high risk interactions.	[13]
OR-MDR	Applied to 42 SNPs in 10 genes related to chronic fatigue syndrome. Both MDR and OR-MDR applied to all possible SNP combinations up to the fourth order.	[15]
MB-MDR	Applied to 282 SNPs in 108 genes of the inflammation pathway of bladder cancer. Eight second-order interactions and 14 third-order interactions were identified.	[34]
GMDR	Applied to 23 SNPs in four genes to identify susceptibility genes for nicotine dependence. GMDR and MDR identified the same interactions.	[16]
GPNN	Applied to 22 SNPs in nuclear-coded mitochondrial complex I genes in a Parkinson's disease cohort. A two-locus interaction between the DLST gene and sex was detected.	[41]
GENN	Applied to 35 SNPs in five genes that encode proteins involved in IL-2/IL-15 signalling. Replicated findings from analysis using MDR.	[18]
RF	Applied to 42 SNPs from the asthma-related ADAM33 gene.	[25]
SNPInterForest	Applied to GWAS data of rheumatoid arthritis from the Wellcome Trust Case Control Consortium (~500 000 SNPs).Two novel interactions identified.	[27]
TRM	Applied to 106 SNPs in six oestrogen receptor-related genes from prostate cancer patients. Interactions identified between SNPs in two genes previously linked to prostate cancer risk and premature ovarian failure.	[33]
RJ	Applied to GWAS data of Crohn's disease containing ~275 000 SNPs. Results validated findings from other GWAS and identified new interactions.	[28]
SVM	Applied to 57 SNPs in 18 genes in a prostate cancer study. Identified high-order interactions between up to five SNPs in line with results from MDR on the same data.	[11]

for which there are already known interactions, there is a pressing need for ML applications that uncover important new interactions in common diseases. No single method has been particularly successful in this respect as yet.

Given the increasingly voluminous genetic data now being produced by next generation sequencing studies, and the emerging evidence that very large numbers of individually low risk variants underlie common diseases, the need for powerful ML models is more pressing than ever. It is evident that current methods require further development before successful application to these enormous data sets can be claimed and their outputs enhance understanding of the genetic epidemiology of disease or become useful in a clinical disease risk predictive setting.

Key Points

- ML methods including multifactor dimensionality reduction, RFs, NNs and SVMs have been developed and applied with some success to detect gene–gene interactions in simulated and disease data.
- Extensions to the original models have facilitated application to large data sets for some approaches but computational and problems of biological interpretation remain.
- Recent two-stage methods which initially reduce the number of potentially interesting SNPs in which to search for interaction may be promising but no single method fully addresses the complexity of the problem.
- Suitable models and applications for the even more complex and voluminous next generation sequencing data are currently lacking.

FUNDING

This work was supported by the Breast Cancer Campaign.

References

1. Moore JH, Williams SM. Epistasis and its implications for personal genetics. *Am J Hum Genet* 2009;**85**:309–20.
2. Hirschhorn JN. Genomewide association studies—illuminating biologic pathways. *N Eng J Med* 2009;**360**:1699701.
3. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 2009;**10**:392–404.
4. Hindorff LA, Sethupathy P, Junkins HA, *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009;**106**:9362–7.
5. Stratton MR, Rahman N. The emerging landscape of breast cancer susceptibility. *Nat Genet* 2008;**40**:17–22.

6. Harlid S, Ivarsson MIL, Butt S, *et al*. Combined effect of low-penetrant SNPs on breast cancer risk. *Br J Cancer* 2012; **106**:389–96.
7. Moore JHP, Ritchie MDP. The challenges of whole-genome approaches to common diseases. *JAMA* 2004;**291**: 1642–3.
8. Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays* 2005;**27**:637–46.
9. Phillips PC. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* 2008;**9**:855–67.
10. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* 2002;**11**:2463–8.
11. Chen S-H, Sun J, Dimitrov L, *et al*. A support vector machine approach for detecting gene-gene interaction. *Genet Epidemiol* 2008;**32**:152–67.
12. Ritchie MD, Hahn LW, Roodi N, *et al*. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001;**69**:138–47.
13. Gui J, Andrew AS, Andrews P, *et al*. A robust multifactor dimensionality reduction method for detecting gene–gene interactions with application to the genetic analysis of bladder cancer susceptibility. *Ann Hum Genet* 2011;**75**:20–8.
14. Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol* 2003;**24**:150–7.
15. Chung Y, Lee SY, Elston RC, *et al*. Odds ratio based multifactor-dimensionality reduction method for detecting gene–gene interactions. *Bioinformatics* 2007;**23**:71–6.
16. Lou X-Y, Chen G-B, Yan L, *et al*. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *Am J Hum Genet* 2007;**80**:1125–37.
17. Motsinger A, Dudek S, Hahn L, *et al*. Comparison of neural network optimization approaches for studies of human genetics. *Lect Notes Comp Sci* 2006;**3907**:103–14.
18. Motsinger-Reif AA, Dudek SM, Hahn LW, *et al*. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genet Epidemiol* 2008;**32**:325–40.
19. Lucek PR, Ott J. Neural network analysis of complex traits. *Genet Epidemiol* 1997;**14**:1101–6.
20. Ritchie MD, Motsinger AA, Bush WS, *et al*. Genetic programming neural networks: a powerful bioinformatics tool for human genetics. *Appl Soft Comput* 2007;**7**:471–9.
21. Koza JR, Rice JP. 'Genetic generation of both the weights and architecture for a neural network'. *Neural Networks, 1991., IJCNN-91-Seattle International Joint Conference* 1991, Vol. 392, pp. 397–404.
22. Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 2010;**26**:445–55.
23. Ritchie M, White B, Parker J, *et al*. Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. *BMC Bioinformatics* 2003;**4**:28.
24. Hu T, Sinnott-Armstrong N, Kiralis J, *et al*. Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics* 2011;**12**:364.
25. Bureau A, Dupuis J, Falls K, *et al*. Identifying SNPs predictive of phenotype using random forests. *Genet Epidemiol* 2005;**28**:171–82.
26. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.
27. Yoshida M, Koike A. SNPInterForest: a new method for detecting epistatic interactions. *BMC Bioinformatics* 2011;**12**: 469.
28. Schwarz DF, Konig IR, Ziegler A. On safari to Random Jungle: a fast implementation of random forests for high-dimensional data. *Bioinformatics* 2010;**26**:1752–8.
29. Liu C, Ackerman H, Carulli J. A genome-wide screen of gene–gene interactions for rheumatoid arthritis susceptibility. *Human Genetics* 2011;**129**:473–85.
30. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**:273–97.
31. Listgarten J, Damaraju S, Poulin B, *et al*. Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. *Clin Cancer Res* 2004;**10**:2725–37.
32. Bi J, Bennett K, Embrechts M, *et al*. Dimensionality reduction via sparse support vector machines. *J Mach Learn Res* 2003;**3**:1229–43.
33. Lin H-Y, Ann Chen Y, Tsai Y-Y, *et al*. TRM: a powerful two-stage machine learning approach for identifying SNP-SNP interactions. *Ann Hum Genet* 2012;**76**:53–62.
34. Calle ML, Urrea V, Vellalta G, *et al*. Improving strategies for detecting genetic patterns of disease susceptibility in association studies. *Stat Med* 2008;**27**:6532–46.
35. Cattaert T, Calle ML, Dudek SM, *et al*. Model-based multifactor dimensionality reduction for detecting epistasis in case–control data in the presence of noise. *Ann Hum Genet* 2011;**75**:78–89.
36. Wan X, Yang C, Yang Q, *et al*. BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am J Hum Genet* 2010;**87**:325–40.
37. De Lobel L, Geurts P, Baele G, *et al*. A screening methodology based on Random Forests to improve the detection of gene–gene interactions. *Eur J Hum Genet* 2010;**18**: 1127–32.
38. Thornton-Wells TA, Moore JH, Haines JL. Genetics, statistics and human disease: analytical retooling for complexity. *Trends Genet* 2004;**20**:640–7.
39. Wang Y, Liu G, Feng M, *et al*. An empirical comparison of several recent epistatic interaction detection methods. *Bioinformatics* 2011;**27**:2936–43.
40. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 2005;**37**:413–7.
41. Motsinger A, Lee S, Mellick G, *et al*. GPNN: Power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. *BMC Bioinformatics* 2006;**7**:39.