BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 24 2007, pages 3280–3288
doi:10.1093/bioinformatics/btm522

Genome analysis

Detecting high-order interactions of single nucleotide polymorphisms using genetic programming

Robin Nunkesser[1,2,*], Thorsten Bernholt[1,2], Holger Schwender[1,3], Katja Ickstadt[1,3] and Ingo Wegener[1,2]

[1]Collaborative Research Center 475, [2]Department of Computer Science and [3]Department of Statistics, University of Dortmund, Dortmund, Germany

Received on July 12, 2007; revised on October 11, 2007; accepted on October 14, 2007

Advance Access publication November 15, 2007

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Not individual single nucleotide polymorphisms (SNPs), but high-order interactions of SNPs are assumed to be responsible for complex diseases such as cancer. Therefore, one of the major goals of genetic association studies concerned with such genotype data is the identification of these high-order interactions. This search is additionally impeded by the fact that these interactions often are only explanatory for a relatively small subgroup of patients. Most of the feature selection methods proposed in the literature, unfortunately, fail at this task, since they can either only identify individual variables or interactions of a low order, or try to find rules that are explanatory for a high percentage of the observations. In this article, we present a procedure based on genetic programming and multi-valued logic that enables the identification of high-order interactions of categorical variables such as SNPs. This method called GPAS cannot only be used for feature selection, but can also be employed for discrimination.

Results: In an application to the genotype data from the GENICA study, an association study concerned with sporadic breast cancer, GPAS is able to identify high-order interactions of SNPs leading to a considerably increased breast cancer risk for different subsets of patients that are not found by other feature selection methods. As an application to a subset of the HapMap data shows, GPAS is not restricted to association studies comprising several 10 SNPs, but can also be employed to analyze whole-genome data.

Availability: Software can be downloaded from http://ls2-www.cs.uni-dortmund.de/~nunkesser/#Software

Contact: robin.nunkesser@uni-dortmund.de

1 INTRODUCTION

Variations in the human genome can alter the risk of developing a disease. The by far most common type of such genetic variations are single nucleotide polymorphisms (SNPs), which occur when at a single base pair position different base alternatives exist. Since a SNP is typically biallelic, it can take three forms: a SNP is of the homozygous reference (or the homozygous variant) genotype if both chromosomes show the base that more (or less) frequently occur in the population, and it is of the heterozygous genotype if one of the bases is the less, and the other is the more frequent alternative.

One of the major goals of association studies is to identify SNPs and SNP interactions that lead to a higher disease risk. Since individual SNPs typically only have a slight to moderate effect—in particular, when considering complex diseases such as sporadic breast cancer—the focus is on the detection of interactions (Culverhouse *et al.*, 2002; Garte, 2001). The search for such interacting SNPs is additionally impeded by the facts that the interactions are usually of a high order, and that they are explanatory for relatively small subgroups of the patients (Pharoah *et al.*, 2004).

Various methods have been suggested for and applied to genotype data to identify SNP interactions. These procedures reach from exhaustive searches based on, e.g. multiple testing approaches (Boulesteix *et al.*, 2007; Goodman *et al.*, 2006; Marchini *et al.*, 2005; Ritchie *et al.*, 2001) to methods based on discrimination procedures (Lunetta *et al.*, 2004). For overviews on such approaches, see Heidema *et al.* (2006) and Hoh and Ott (2003).

One of the most promising methods is logic regression (Ruczinski *et al.*, 2003), an adaptive classification and regression procedure that tries to identify Boolean combinations of binary variables associated with the response (e.g. the case-control status). In several comparisons with other regression or discrimination approaches, logic regression has shown a good performance in its application to SNP data (Kooperberg *et al.*, 2001; Ruczinski *et al.*, 2004; Schwender, 2007; Witte and Fijal, 2001). Moreover, logic regression can be employed for detecting interactions and quantifying their importance (Kooperberg and Ruczinski, 2005; Schwender and Ickstadt, 2007).

For an application of logic regression to genotype data, each SNP needs to be coded by (at least) two dummy variables, as logic regression can only handle binary predictors, but SNPs can take three forms. Although this coding can be done in a biologically meaningful way (one dummy variable codes for a dominant effect, and the other for a recessive effect), it might be preferable to include each SNP as one variable in the analysis. Furthermore, the logic expressions generated by logic regression should be transformed into a disjunctive normal form

*To whom correspondence should be addressed.

(DNF) to identify the interactions, as the monomials included in the DNF can be interpreted as interactions.

Therefore, our procedure called GPAS (Genetic Programming for Association Studies) employs multi-valued logic, and attempts to detect DNFs associated with the response directly. To search for such DNFs, genetic programming (Koza, 1993) is used. Genetic programming naturally provides not a single best model, but a set of models (called individuals) that fit almost equally well, which is an advantage in the analysis of genotype data in which many competing models might exist.

In the following section, GPAS is introduced in detail. Afterwards, GPAS is applied to the genotype data from the GENICA study, a study dedicated to the identification of genetic and gene-environment interactions leading to a higher risk of developing sporadic breast cancer. In the analysis of this data set, GPAS is able to detect high-order SNP interactions associated with the case-control status. But GPAS is not restricted to association studies comprising several 10 SNPs. It can also be used to analyze data from whole-genome studies. To exemplify this, GPAS is also applied to a subset of the HapMap data (The International HapMap Consortium, 2003). Moreover, GPAS cannot only be employed for feature selection, but also for discrimination. In a comparison with other discrimination methods, GPAS shows the smallest misclassification rates when applied both to the real data sets from the HapMap and the GENICA study and to simulated data.

2 METHODS

We propose to use evolutionary algorithms—more precisely genetic programming (Koza, 1993)—for the analysis of genotype data.

In genetic programming, a set of *individuals* called *population* undergoes adaptions and afterwards a selection process based on *fitness* leading to a new *generation* of individuals. This procedure summarized in Algorithm 1 is iterated until a *termination criterion* is fulfilled.

ALGORITHM 1 (Basic Genetic Programming Algorithm).

(1) Create an initial random population.
(2) Perform the following steps on the current generation:
   (a) Select individuals in the population based on a selection scheme.
   (b) Adapt the selected individuals.
   (c) Evaluate the fitness value of the adapted individuals.
   (d) Select individuals for the next generation according to a selection scheme.
(3) If the termination criterion is fulfilled, then output the final population. Otherwise, set the next generation as current and go to step 2.

2.1 Genetic programming for association studies

In the following, we customize the basic genetic programming algorithm presented in Algorithm 1 for our purpose, leading to our method GPAS.

2.1.1 Structure of the individuals In GPAS, multi-valued logic expressions in disjunctive normal form (DNF) are used as the structure for the individuals, where these logic expressions may exhibit any number of input states. In the application to SNP data, e.g. an input can take one of the following three states: (1) coding for the homozygous reference, (2) heterozygous and (3) homozygous variant.

A logic expression in DNF is a disjunction of one or more monomials, where a monomial is a single literal or a conjunction of literals. Given, e.g. a set of variables $X_1, \ldots, X_m$, each of which can take K values, the literals used in GPAS are

$$(X_i = k) \text{ and } (X_i \neq k), \quad k = 1, \ldots, K, \ i = 1, \ldots, m.$$

In Figure 1, examples of generic tree representations of such logic expressions in DNF resulting from analyzing SNPs are shown. For example, the tree labeled 'Original individual' represents the logic expression

$$L = \big((\mathrm{SNP}_1 = 3)\big) \ \vee \ \big((\mathrm{SNP}_2 \neq 1) \wedge (\mathrm{SNP}_3 = 1)\big).$$

When used as a predictor in a case-control study, a patient would be classified as case if L is true, i.e. if all SNPs in at least one of the two monomials $((\mathrm{SNP}_1 = 3))$ and $\big((\mathrm{SNP}_2 \neq 1) \wedge (\mathrm{SNP}_3 = 1)\big)$ show the genotypes indicated by the corresponding literals. Otherwise, the patient would be classified as control.

To store a logic expression in memory we use *trees* (Cormen *et al.*, 2001) that are built according to the depicted tree representation as data structure. Using trees allows some very flexible and inexpensive operations: all of the adaptions described in the following are possible in amortized constant time when the children of a node in the tree are stored in a dynamic array.

2.1.2 Operations for adapt individuals Initially, a population composed of two individuals, each consisting of one randomly selected literal, is created (corresponding to step 1 of Algorithm 1).

The set of candidate individuals for a new generation is constructed in steps 2a and b by selecting

- all individuals for *reproduction*, i.e. copying all individuals from the current generation,
- two individuals uniformly at random for *crossover*, i.e. combining one of the two individuals with one randomly chosen monomial from the other individual to create a new individual,
- five individuals uniformly at random for *mutation*, i.e. applying a random change to each of the individuals, where each of the following possible mutations is applied to exactly one of the five individuals:
  - inserting a new literal,
  - deleting a literal,
  - replacing a literal by a new literal,
  - inserting a new literal as a new monomial,
  - deleting a monomial.

In the latter adaption, the literals or monomials that should be deleted are chosen uniformly at random, and the new literals are also selected at random and inserted into a randomly chosen monomial or as a new monomial. An overview on the crossover and the mutations is given in Figure 1.

Note that the usage of crossover is discussed controversial (see e.g. Banzhaf *et al.*, 1998). However, the crossover operation we propose does not disrupt the structure of the individuals and is therefore different from the criticized crossover operations. In the applications considered in this article, it accelerates computation.

2.1.3 Fitness and selection To determine which of the new and reproduced individuals are selected to be part of the next generation, we compute the fitness for each individual and select the best ones (corresponding to steps 2c and d of Algorithm 1, respectively).

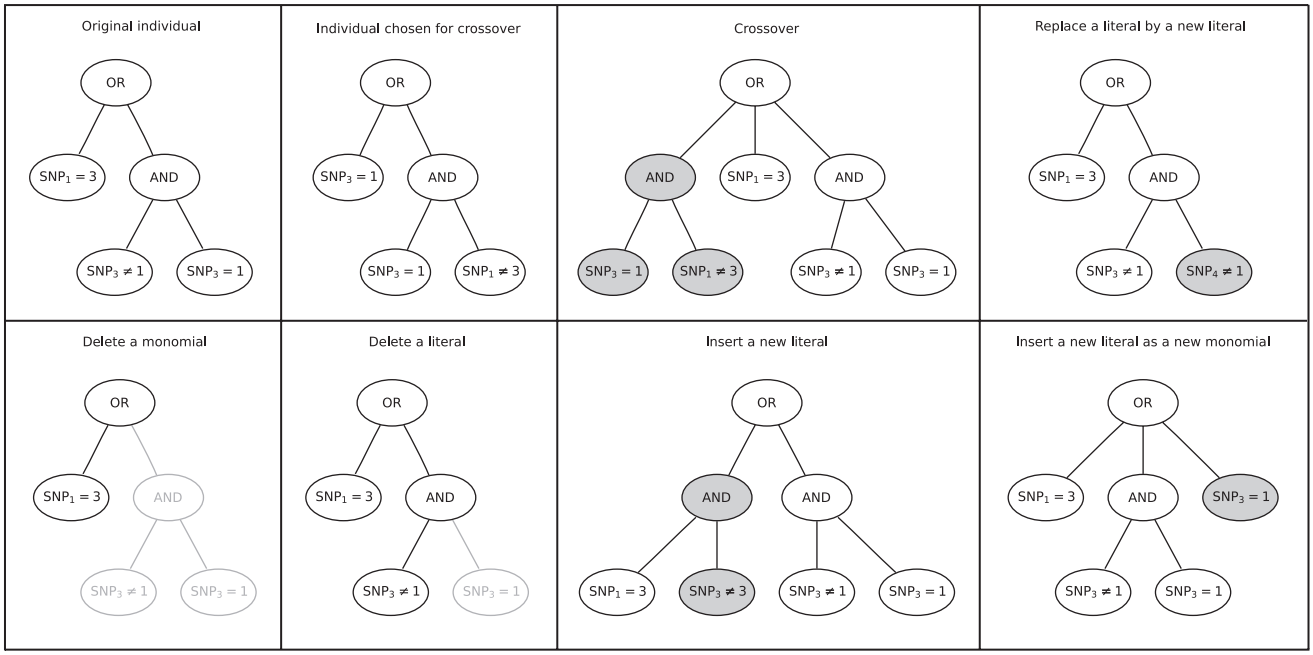

Fig. 1. Examples for the crossover and the different mutations used in GPAS.

We are interested in logic expressions that explain as many observations as possible, while being as short as possible. To achieve both goals equally well, we conduct a multi-objective optimization by using multidimensional fitness values. The basic objectives of our optimization may be transferred to fitness values easily. Explaining as many observations as possible, for instance, translates to fitness values measuring the amount of data values fulfilled.

In the context of multi-objective optimization, an individual *dominates* another individual, if at least one component of its fitness value is superior, and none is inferior. An individual is *pareto optimal*, if it is not dominated by another individual. Consequently, we seek to find pareto optimal individuals that offer a set of well fitting models.

For the new generation of individuals that is derived after the adaptions, we choose only individuals that are not dominated by other individuals. Thus, we conduct a *domination selection*. As a consequence, none of the objectives is preferred, i.e. no additional weighting scheme is employed. For our purposes, we use three objectives (see Sections 2.2 and 2.3). That leads to a bigger population and allows a more specialized search. For our two tasks—identification of interesting interactions and discrimination—we employ two slightly different fitness functions that are also described in Sections 2.2 and 2.3, respectively.

The major computational part of the fitness evaluation is to determine the number of cases and controls classified correctly by the logic expression.

For fast fitness computation, we additionally store a bitset in each node of the tree representing the logic expression. The bitset consists of as many bits as there are observations in the data set, and the ith bit is true if the logic expression is true for the SNP forms of the ith observation and false otherwise.

The bitsets of the literals are initially computed for all possible literals. If a monomial of the logical expression is changed during a mutation operation the bitset of the monomial is recomputed using the bitsets of its literals. The computation is sped up, since the bitsets of the other monomials remain unchanged and can be reused to compute the bitset of the whole logic expression. In addition, bitsets are compact and allow fast logic operations. For example, one logic operation of the bitset of the whole logic expression with the bitset describing the case-control status suffices to compute the number of cases and controls predicted correctly.

2.1.4 Termination criterion We need termination criteria for the genetic programming process in order to derive a final population building the models (step 3 of Algorithm 1). Natural termination criteria used by GPAS are the excess of a certain number of generations or of a certain fitness value. Another possibility is to terminate the execution if the algorithm *stagnates*, i.e. no new individuals survived selection for a given number of generations.

2.2 Identification of interactions

A major influence factor on the objective of an analysis is the choice of the fitness evaluation function. Interactions that explain subsets of the cases have to contradict with as few controls as possible. We, therefore, employ a fitness evaluation function that emphasizes this by including the number of correctly predicted controls in two of the objectives.

The fitness of an individual is thus evaluated by the fitness function f_1 that maps a logic expression to the following triple (corresponding to three objectives):

- (Maximize the) mean of the proportions of correctly classified cases and correctly classified controls.
- (Maximize the) number of controls the logic expression correctly predicts.
- (Minimize the) length of the logic expression, i.e. the number of literals of the logic expression.

Including the number of correctly predicted controls in two objectives leads to a preference of models that contradict with few controls during the domination selection. After the search, we obtain a population of individuals that are not dominated by each other.

A further modification of the general genetic algorithm is that we do not allow individuals to become too big. In the search for high-order

interactions, we, furthermore, prohibit the algorithm from constructing polynomials, i.e. individuals, with more than two monomials.

To aid the detection of high-order interactions, we additionally devise a visualization of the resulting models. The interactions in the model are displayed in a tree showing many different interactions at a glance. To obtain this visualization (for an example of a resulting tree, see Fig. 2), we proceed as follows:

(1) Obtain the set $\mathcal{M}$ of all monomials occurring in the resulting models.
(2) Search for the most common literal ℓ in $\mathcal{M}$, and determine the set $\mathcal{M}_\ell$ of monomials containing ℓ.
(3) Exclude ℓ from all monomials in $\mathcal{M}_\ell$ to construct $\mathcal{M}_{-\ell}$.
(4) Repeat steps 2–3 with $\mathcal{M} := \mathcal{M}_{-\ell}$ and $\mathcal{M} := \mathcal{M} \backslash \mathcal{M}_\ell$ until $\mathcal{M} = \emptyset$.

We additionally store information on how often the resulting interactions and partial interactions occur, and on how many observations they explain.

2.3 Discrimination

In the case of discrimination, the correct prediction of cases and controls is treated as equally important. Therefore, the first objective of f_1 is replaced by (maximize the) number of cases the logic expression predicts correctly leading to the fitness function f_2. Thus, predicting cases is treated in the same way as predicting controls. To elucidate the difference between f_1 and f_2 consider, e.g. two individuals a and b with the same length, where a predicts 50% of the cases and 90% of the controls correctly and b predicts 89% of the cases and 50% of the controls correctly. When we use f_1, a dominates b, while it does not dominate b when we use f_2.

Additionally, we restrict the size of the individuals, but not the number of monomials comprised in an individual.

For class prediction of new observations, either the single best individual, i.e. the individual with the lowest misclassification rate (for a given length), is used, or an ensemble of models is considered either by averaging over a set consisting of the best individuals or by applying bagging (Breiman, 1996) to GPAS.

2.4 GPAS

To summarize, we propose the following specialized genetic programming algorithm called GPAS for the analysis of genotype data.

ALGORITHM (GPAS).

(1) Create an initial random population composed of two individuals each of which consists of one randomly selected literal.
(2) Perform the following steps on the current generation:
   (a) Select all individuals in the population for reproduction, and draw seven of the individuals uniformly at random.
   (b) Conduct each of the following adaptions to one (mutations) or two (crossover) of the seven randomly selected individuals.
      - Perform a crossover.
      - Insert a new literal.
      - Delete a literal.
      - Replace a literal by a new literal.
      - Insert a new literal as a new monomial.
      - Delete a monomial.
   (c) Evaluate the fitness value of the adapted and reproduced individuals with fitness function f_1 or f_2.
   (d) Select all adapted and reproduced individuals that are not dominated for the next generation.
(3) If the termination criterion is fulfilled, then output the final population. Otherwise, set the next generation as current and go to step 2.

3 DATA SETS

3.1 GENICA

The GENICA study is an age-matched and population-based case-control study carried out by the Interdisciplinary Study Group on Gene ENvironment Interaction and Breast CAncer in Germany (http://www.genica.de), a joint initiative of researchers dedicated to the identification of genetic and environmental factors associated with sporadic breast cancer. Cases and controls have been recruited in the greater Bonn, Germany, region. Apart from exogenous risk factors such as reproduction variables, hormone variables and life style factors, the genotypes of about 100 polymorphisms have been assessed from these women (for details on the GENICA study, see Justenhoven *et al.*, 2004).

In this article, the focus is on a subset of the genotype data from the GENICA study. More precisely, data of 1258 women (609 cases and 649 controls) and 63 SNPs are available for the analysis. Since a small number of observations show a large number of missing values, we remove all women with more than five missing values leading to a total of 1191 observations (561 cases and 630 controls). The remaining missing values are replaced SNP-wise by random draws from the marginal distribution.

3.2 HapMap

The goals of the International HapMap Project (The International HapMap Consortium, 2003; http://www.hapmap.org) are the development of a haplotype map of the human genome and the comparison of genetic variations of individuals from different populations. To achieve this goal, millions of SNPs have been genotyped for each of 270 people from four different populations.

In this article, the SNP data of 45 unrelated Han Chinese from Beijing and 45 unrelated Japanese from Tokyo measured by employing the Affymetrix GeneChip Mapping 500 K Array Set are considered.

This array set consists of two chips (the Nsp and the Sty array named after the restriction enzymes used on these chips) each enabling the genotyping of about 250 000 SNPs. Here, we focus on the BRLMM genotypes (Bayesian Robust Linear Model with Mahalanobis distance; Affymetrix 2006) of the 262 264 SNPs from the Nsp array that can be downloaded from http://www.affymetrix.com/support/technical/sample_data/500k_hapmap_genotype_data.affx. All SNPs showing one or more missing genotypes (54 400 SNPs), for which not all three genotypes are observed (75 481 SNPs), or that have a minor allele frequency less than or equal to 0.1 (10 609 SNPs) are excluded in this order from the analysis leading to a data set composed of the genotypes of 121 774 SNPs and 90 individuals.

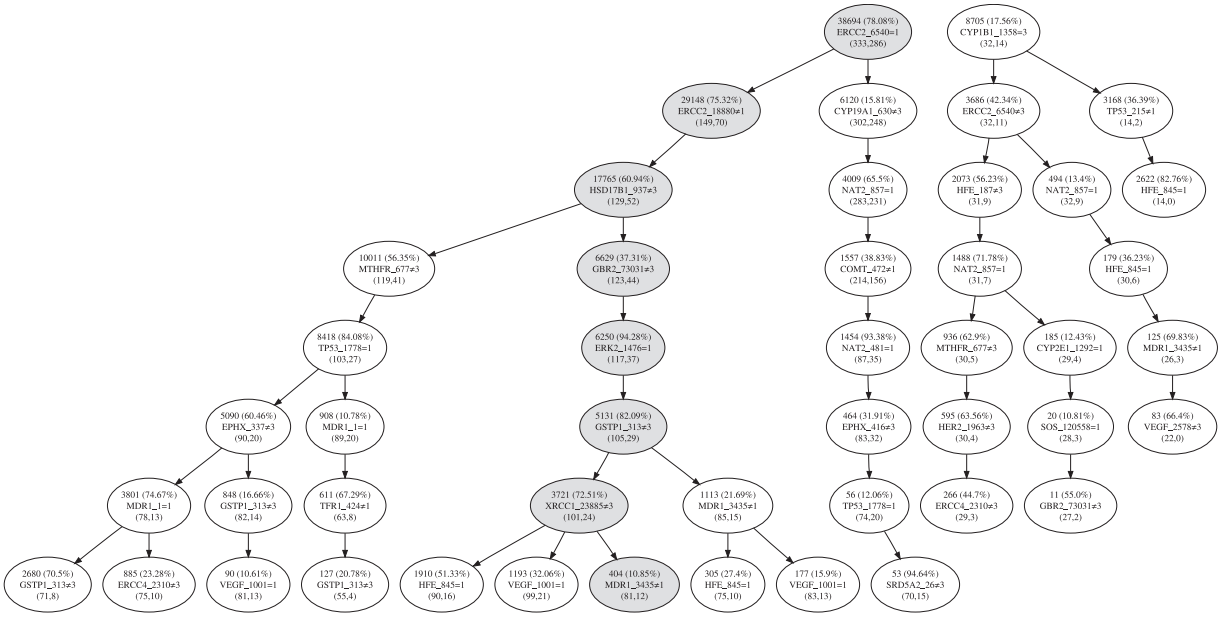

Fig. 2. Excerpt from the tree visualization of the models resulting from the application of GPAS to the GENICA data set. Each path from the root to an inner node or leaf represents an interaction occurring in the final population. The first line in each node consists of the number of monomials containing the corresponding interaction and the percentage of monomials consisting of the ancestral interaction that also contain the literal represented by the node, where this literal is displayed in the second line. The third line shows the number of cases and controls explained by the corresponding interaction.

3.3 Simulated data

In the discrimination case, simulated data mimicking SNP data from real association studies are considered as well. Genotypes for 50 SNPs and 1000 observations are randomly generated, where each of these unlinked SNPs exhibits a minor allele frequency of 0.25. The case-control status y is then randomly drawn from a Bernoulli distribution with mean $\text{Prob}(Y=1)$, where

$$\text{logit}\left(\text{Prob}(Y=1)\right) = -0.5 + 1.5L_1 + 1.5L_2$$

with $L_1 = (\text{SNP}_3 = 1) \wedge (\text{SNP}_9 = 1) \wedge (\text{SNP}_{10} = 0)$ and $L_2 = (\text{SNP}_6 \neq 1) \wedge (\text{SNP}_7 = 1)$ such that the probability for being a case is 0.924 if for an observation both L_1 and L_2 are true, and is 0.731 if one of these logic expressions is true. This probability is still 0.378 if an observation does not exhibit one of these two interactions intended to be influential for the risk of developing the disease of interest in this simulated association study. A reason for this might be that there are other genetic (or environmental) factors that have not been considered in this study, but have an influence on the disease risk.

This procedure is repeated 50 times such that 50 data sets are generated.

4 RESULTS

The following analyses are conducted on a Pentium 4 CPU with 2.56 GHz and 1024 MB of RAM.

4.1 Identification of interesting SNP interactions

In association studies concerned with sporadic breast cancer, it is assumed that not individual SNPs, but combinations of many SNPs have an high impact on the cancer risk, and that each of these interactions is a risk factor for a particular (relatively small) subgroup of patients (Pharoah *et al.*, 2004). In the analysis of the GENICA data set, we are thus interested in identifying high-order interactions explaining several 10 cases, but only a few controls.

As mentioned in Section 2, we therefore constrain each individual in GPAS to consist of a maximum of two monomials. As $\text{SNP}_i \neq 1$ codes for a dominant effect of SNP_i, and $\text{SNP}_i = 3$ for a recessive effect, we restrict the set of literals used in GPAS to these two literals and their respective complements, i.e. $\text{SNP}_i = 1$ and $\text{SNP}_i \neq 3$.

In this application of GPAS to the GENICA data set, we gather the individuals of 50 independent runs each of which stops after 500 000 generations, which takes ~10 min. From the resulting 49 564 individuals, the tree visualization described in Section 2.2 is constructed. An excerpt from this tree is shown in Figure 2. For example, the eight literals marked by a gray background form an interaction that explains, i.e. a monomial that is true for, 81 cases and only 12 controls, and is contained in 404 of the individuals.

Figure 2 also reveals that the interesting SNP interactions contain $(\text{ERCC2\_6540} = 1) \wedge (\text{ERCC2\_18880} \neq 1)$, i.e. an interaction of the two SNPs ERCC2_6540 (refSNP ID: rs1799793) and ERCC2_18880 (rs1052559) from the gene ERCC2 (Excision Repair Cross-Complementing group 2;

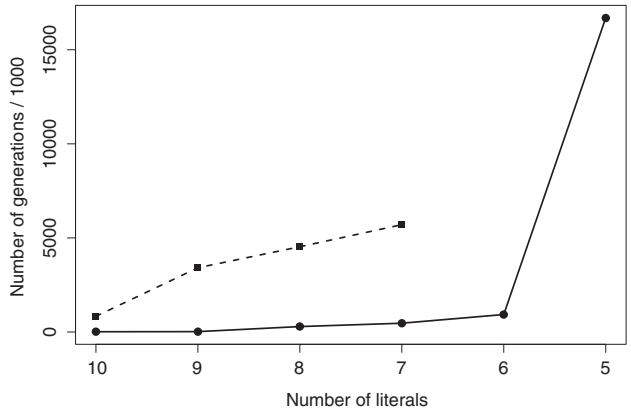

Fig. 3. Number of generations (in thousands) in which individuals of certain lengths predicting all observations correctly are found in the application of GPAS to the HapMap data set using the real ethnicity (solid line) and a random group assignment (dashed line).

formerly XPD), which itself explains 149 cases and 70 controls. This two-way interaction has already been found by Justenhoven *et al.* (2004) and by Schwender and Ickstadt, (2007), but they were not able to identify interactions of higher orders with better odds ratios.

For comparison, GPAS is again applied to the GENICA data set using random assignments of the case-control status to the women. In this case, all detected individuals show ratios of explained cases to explained controls that are smaller than the ratios of comparable interactions found in the original application. For example, the individual that is best comparable with the interaction that is marked by gray background in Figure 2 and explains 81 cases and 12 controls is an logic expression that is true for 89 cases and 30 controls.

To examine if the exclusion of $(\text{SNP}_i = 2)$ and $(\text{SNP}_i \neq 2)$ has a large influence on the detection of interesting interactions, we also apply GPAS to the GENICA data set using the complete set of literals. In this analysis, some of the literals in the identified monomials are indeed of this type. However, these literals have mostly only a small effect, or they are equivalent to, e.g. $(\text{SNP}_i = 1)$. For example, the interaction

$$(\text{ERCC2\_6540} = 1) \wedge (\text{ERCC2\_18880} \neq 1) \wedge (\text{TFR\_424} \neq 1) \wedge (\text{CYP1A1\_2452} = 1) \wedge (\text{MDR1\_1} \neq 2) \wedge (\text{TP53\_1778} \neq 2)$$

detected in this application, which explains, i.e. is true for, 73 cases and 16 controls, contains two literals of the form $(\text{SNP}_i = 2)$. However, $(\text{MDR1\_1} \neq 2)$ is actually $(\text{MDR1\_1} = 1)$, as none of the observations exhibit the homozygous variant genotype at this SNP, and replacing $(\text{TP53\_1778} \neq 2)$ by $(\text{TP53\_1778} = 1)$ would reduce the number of correctly predicted cases from 73 to 72, while the number of explained controls stays at 16.

To exemplify that GPAS is not restricted to data sets consisting of several ten to a few hundred SNPs, but can also be applied to data from whole -genome studies, we apply GPAS to the subset of the HapMap data set described in Section 3.2.

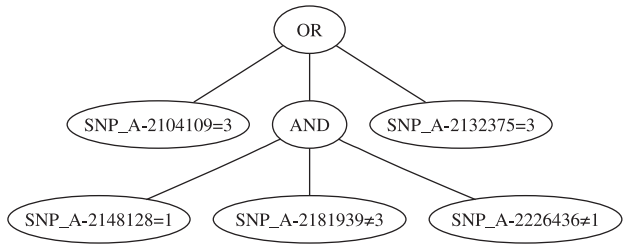

Fig. 4. Individual composed of five SNPs that is identified by GPAS[] in the HapMap data set. It can be used to distinguish between Japanese and Han Chinese.

As it might be possible that individual SNPs have a large influence in this example, we do not restrict the number of monomials in an individual. Furthermore, we only run GPAS once but without a termination criterion. All other settings remain unchanged compared to the analysis of the GENICA data set.

After running for 9 min, GPAS detects an individual composed of 10 literals in generation 13 683 that can be used to distinguish between the Japanese and the Han Chinese unambiguously: if at least one of the six monomials $\big((\text{SNP\_A} - 1840639 = 1)\big)$, $\big((\text{SNP\_A} - 1862578 = 1)\big)$, $\big((\text{SNP\_A} - 1888933 = 3)\big)$, $\big((\text{SNP\_A} - 1983282 = 1) \wedge (\text{SNP\_A} - 2227333 = 3)\big)$, $\big((\text{SNP\_A} - 1849099 \neq 1) \wedge (\text{SNP\_A} - 2046537 \neq 1)\big)$, and $\big((\text{SNP\_A} - 2030395 = 1) \wedge (\text{SNP\_A} - 1940113 \neq 1) \wedge (\text{SNP\_A} - 4200881 \neq 3)\big)$ is true, then the person is from Japan (or more exactly, from Tokyo). Otherwise, it is a Han Chinese from Beijing.

This individual can still be optimized by reducing the number of SNPs (which is the third objective used in GPAS). Shortly after detecting this individual, GPAS finds individuals down to length six (see Fig. 3), and finally in generation 16 691 641 an individual composed of five literals/SNPs and displayed in Figure 4 is identified, where each of these individuals predict all observations correctly.

For comparison, GPAS is applied to the HapMap data set using random group assignments. Not surprisingly, these applications also lead to perfect separations of the two groups. However, the detected logic expressions are composed of more than five individuals, and it takes much longer to detect these individuals (for an example of the results of such an application, see Fig. 3).

4.2 Discrimination

To examine how the misclassification rate depends on the number of variables in the model, GPAS is applied to the GENICA data set considering individuals composed of differing numbers of literals. For each number of variables considered, we let GPAS run for 10 000 generations, which takes about 1 min for each run.

For comparison, the GENICA data set is also analyzed using logic regression (Ruczinski *et al.*, 2003), where the number of variables allowed is constrained in the different applications. Since the cases and the controls are age-matched in the GENICA study, conditional logistic regression (Breslow and

Day, 1980) that takes this matching into account is also applied to this data set. Here, the variables are chosen by forward selection.

As both logic regression and conditional logistic regression require binary predictors, the ith SNP, $i = 1, \ldots, m$, is split into the two dummy variables

SNP_{i1}: 'SNP$_i$ is not of the homozygous reference genotype.'

SNP_{i2}: 'SNP$_i$ is of the homozygous variant genotype.'

where SNP_{i1} codes for a dominant effect of SNP$_i$, and SNP_{i2} for a recessive effect. Note that SNP_{i1}, $\overline{\text{SNP}_{i1}}$, SNP_{i2} and $\overline{\text{SNP}_{i2}}$ correspond to $\text{SNP}_i \neq 1$, $\text{SNP}_i = 1$, $\text{SNP}_i = 3$ and $\text{SNP}_i \neq 3$, respectively.

In Figure 5, the resulting misclassification rates estimated by 10-fold cross-validation are displayed. This figure shows that the misclassification rates of both GPAS and logic regression are equal if the number of literals is less than 3. This is due to the fact that both use $((\text{ERCC2\_6540} = 1))$ or $((\text{ERCC2\_6540} = 1) \wedge (\text{ERCC2\_18880} \neq 1))$, respectively, as classification rule in any of the respective iterations of the cross-validation. However, the misclassification rate of GPAS becomes smaller than the one of logic regression if the models are allowed to be composed of three to eight variables. Both GPAS and logic regression outperform conditional logistic regression that exhibits a minimum estimated misclassification rate of 42.4%. As in the applications of GPAS and logic regression, ERCC2_6540_1 and ERCC2_18880_1 are always identified by the conditional logistic regression to be the two most important variables. However, when considering models composed of these two variables, the misclassification rate of conditional logistic regression is higher than the one of the other two approaches (see Fig. 5). A reason for this is that in the conditional logistic regression models ERCC2_6540_1 and ERCC2_18880_1 are considered as additional effects, whereas in GPAS and logic regression the interaction $\big((\text{ERCC2\_6540} = 1) \wedge (\text{ERCC2\_18880} \neq 1)\big)$ is used as predictor. This shows an advantage of GPAS (and logic regression): while in approaches such as conditional logistic regression it is necessary to include all $\binom{n}{p}$ p-way interactions of n variables to identify important p-way interactions, GPAS is able to detect such interactions using only the n variables as inputs.

For a comparison of GPAS with further tree-based discrimination methods, CART (Breiman *et al.*, 1984), bagging (Breiman, 1996) and Random Forests (Breiman, 2001) are applied to the GENICA data set, where the parameters of the latter two procedures are optimized over several values. (In both bagging and Random Forests, different numbers of trees are considered. Additionally, different numbers of randomly chosen variables at each node are used in Random Forests.)

In Table 1, the misclassification rates of these applications are summarized. This table reveals that GPAS leads to less misclassifications than the other discrimination procedures.

For the application of these discrimination methods to the HapMap data set, the number of variables has to be reduced to a size that these approaches can handle. We therefore use the Significance Analysis of Microarrays (SAM; Tusher *et al.*, 2001) adapted for categorical data (Schwender, 2005) to reduce

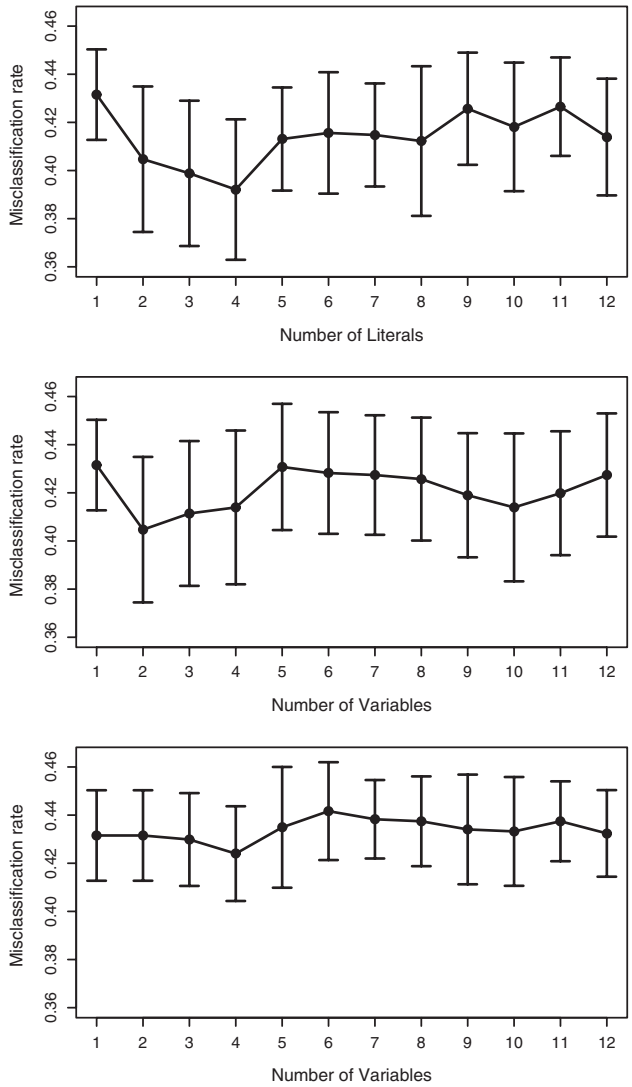

Fig. 5. Misclassification rates of GPAS (top), logic regression (middle) and conditional logistic regression (bottom) in their applications to the GENICA data set with restricted numbers of literals/variables in the individuals/models. While the solid dots mark the respective mean misclassification rate, the vertical lines represent the corresponding 95% confidence intervals.

the number of SNPs from 121 774 to 157, where this subset of SNPs exhibits an estimated FDR (False Discovery Rate) of 0.069.

All discrimination methods are then applied to this subset of SNPs, and the misclassification is estimated by 9-fold cross-validation, where each of the nine subsets is composed of five randomly chosen Han Chinese and five randomly chosen Japanese.

Since for each of the training sets several models might exist that predict all training observations correctly, we use the bagging version of GPAS to stabilize the discrimination. We also stop after 10 000 generations, which takes ~12 min for one training (consisting of 100 runs due to the use of bagging).

Table 1. Means and SDs of the misclassification rates of the applications of several discrimination methods to the GENICA, the HapMap and the simulated data sets

		GENICA	HapMap	Simulation
GPAS	Mean	0.392	0.011	0.335
	SD	0.047	0.034	0.025
Logic Regression	Mean	0.405	0.144	0.342
	SD	0.049	0.103	0.022
CART	Mean	0.429	0.356	0.371
	SD	0.034	0.101	0.025
Bagging	Mean	0.453	0.022	0.382
	SD	0.031	0.044	0.018
Random Forests	Mean	0.450	0.011	0.379
	SD	0.021	0.034	0.018

As Table 1 shows, both GPAS and Random Forests only misclassify one observation, whereas the discrimination methods that use a single model as classification rule, i.e. CART and logic regression, show a comparatively high misclassification rate.

Furthermore, the five discrimination methods are applied to the 50 simulated data sets described in Section 3.3, where each of these data sets is used once as training set and once as test set. (The classification rule trained on data set 1 is tested on data set 2, the rule trained on data set 2 is tested on data set 3, and so on.) As Table 1 reveals, GPAS again shows a misclassification rate that is smaller than the ones of the four other discrimination procedures, and that comes close to the actual misclassification rate of 32.6%.

5 DISCUSSION

A major goal of association studies is the identification of SNPs and more importantly SNP interactions that lead to a higher risk of developing a disease. When considering complex diseases such as sporadic breast cancer, such interactions are typically of a high order and only explain relatively small subsets of the patients. Thus, approaches are needed that are able to detect these risk factors.

In this article, we have presented a procedure based on genetic programming that can cope with this task. Genetic programming has the advantage that it is a general purpose method that can handle changing demands flexibly such as different fitness functions or size-constraints. In addition, the maintenance of candidate solutions is expedient for the multi-objective problems we tackle.

In the analysis of the GENICA data set, the presented method called GPAS identifies high-order interactions that explain sets of about 100 observations from which only a few are controls. Some of the detected interactions will have an impact on the risk of developing sporadic breast cancer, whereas others will only show up in the considered data set. Thus, the identified interactions need to be tested on an independent data set consisting of new observations to determine which of these interactions are in fact important risk factors.

As the application to the 121 774 SNPs from the HapMap data set shows, GPAS can also be used to analyze whole-genome data. In this application, a logic expression composed of five SNPs is identified by GPAS that allows to unambiguously distinguish between the two HapMap populations Japanese from Tokyo and Han Chinese from Beijing. However, it might be harder to detect influential high-order SNP interactions when the disease status instead of ethnicity is the covariate of interest – in particular if only a few SNPs that themselves might only have a low effect are assumed to be involved in the development of the disease, and the sample size is small.

GPAS is not restricted to feature selection, but can also be employed for classification, where it outperforms other tree-based discrimination methods in the applications to both simulated data and the real data sets from the GENICA and the HapMap study.

Although GPAS has been developed in the context of SNP data, it can also be used to analyze other types of categorical data, where the numbers of levels the variables can take might differ between variables. For example, it can also be applied to non-biallelic genetic polymorphisms such as microsatellites or to haplotypes. However, GPAS currently is not able to take into account the uncertainties that show up if the haplotypes are estimated using procedures such as PHASE (Stephens and Donelly, 2003).

Furthermore, the design of GPAS is flexible: by default, the set of literals is composed of all possible values for any of the variables and their corresponding complements. It is, however, possible to constrain this set of literals. For ordinal data, $>$ and $<$ can be used as operators additionally to or instead of $=$ and $\neq$. Another possibility is to exclude any of the moves. For example, removing crossover from the move set might not worsen the results, but is likely to increase the computation time, as more generations have to be considered before the best solution is found.

Currently, the inputs, but not the output of GPAS can be multi-valued, as we are mainly interested in case-control studies. However, an extension of the two-class to the multi-class case is planned.

Another idea is to formulate—similar to logic regression—GPAS in a regression framework so that continuous responses, that are, e.g. of interest in QTL (Quantitative Trait Loci) analyses, can also be considered.

ACKNOWLEDGEMENTS

Financial support of the Deutsche Forschungsgemeinschaft (SFB 475, 'Reduction of Complexity in Multivariate Data Structures') is gratefully acknowledged. The authors would also like to thank Roland Friend and all partners within the GENICA research network for their cooperation, and in particularly Hermann M. Bolt, Hiltrud Brauch, Ute Hamann, Jan Hengstler, Christina Justenhoven, Sylvia Rabstein and Anne Spickenheuer for helpful discussions and Melanie Schmidt for her help in implementing GPAS.

Conflict of Interest: none declared.

REFERENCES

Affymetrix (2006) BRLMM: an improved genotype calling method for the GeneChip Human Mapping 500k array set. *Technical report*. Affymetrix, Santa Clara, CA.

Banzhaf,W. *et al*. (1998) *Genetic Programming: an Introduction: on the Automatic Evolution of Computer Programs and Its Applications*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Boulesteix,A.L. *et al*. (2007) Multiple testing for SNP-SNP interactions: a flexible asymptotic framework. *Technical report, Sylvia Lawry Centre*. Munich, Germany.

Breiman,L. *et al*. (1984) *Classification and regression trees*. Wadsworth, Belmont, CA.

Breiman,L. (1996) Bagging predictors. *Mach. Learn*., **26**, 123–140.

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

Breiman,L., Friedman,J.H., Olshen,R.A. and Stone,C.J. (1984) *Classification and Regression Trees*. Wadsworth, Belmont, C.A.

Breslow,N.E. and Day,N.E. (1980) *Statistical Methods in Cancer Research: The Analysis of Case-control Studies*. IARC scientific publications, Lyon.

Cormen,T.H. *et al*. (2001) *Introduction to Algorithms*. 2nd edn. The MIT Press, Cambridge, MA.

Culverhouse,R. *et al*. (2002) A perspective on epistasis: limits of models displaying no main effect. *Am. J. Hum. Genet*., **70**, 461–471.

Garte,S. (2001) Metabolic susceptibility genes as cancer risk factors: time for a reassessment? *Cancer Epidemiol. Biomarkers Prev*., **10**, 1233–1237.

Goodman,J.E. *et al*. (2006) Exploring SNP-SNP interactions and colon cancer risk using polymorphism interaction analysis. *Int. J. Cancer*, **118**, 1790–1797.

Heidema,G.A. *et al*. (2006) The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. *Biomed. Genet*., **7**.

Hoh,J. and Ott,J. (2003) Mathematical multi-locus approaches to localizing complex human trait genes. *Nat. Rev. Genet*., **4**, 701–709.

Justenhoven,C. *et al*. (2004) ERCC2 genotypes and a corresponding haplotype are linked with breast cancer risk in a German population. *Cancer Epidemiol. Biomarkers Prev*., **13**, 2059–2064.

Kooperberg,C. and Ruczinski,I. (2005) Identifying interacting SNPs using Monte Carlo logic regression. *Genet. Epidemiol*., **28**, 157–170.

Kooperberg,C. *et al*. (2001) Sequence analysis using logic regression. *Genet. Epidemiol*., **21**, 626–631.

Koza,J.R. (1993) *Genetic Programming – On the Programming of Computers by Means of Natural Selection*. The MIT Press, Cambridge, MA.

Lunetta,K. L., Hayward,L. B., Segal,J. and van Eerdewegh,P. (2004) Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet*, **10**.

Marchini,J. *et al*. (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet*., **37**, 413–416.

Pharoah,P.D. *et al*. (2004) Association studies for finding cancer-susceptibility genetic variants. *Nat. Rev. Cancer*, **4**, 850–860.

Ritchie,M.D. *et al*. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet*., **69**, 138–147.

Ruczinski,I. *et al*. (2003) Logic regression. *J. Comput. Graph. Stat*., **12**, 475–511.

Ruczinski,I. *et al*. (2004) Exploring interactions in high-dimensional genomic data: an overview of logic regression, with applications. *J. Mult. Anal*., **90**, 178–195.

Schwender,H. (2005) Modifying microarray analysis methods for categorical data – SAM and PAM for SNPs.Weihs,C. and Gaul,W. (eds.) *Classification – The Ubiquitous Challenge*. Springer, Heidelberg, pp. 370–377.

Schwender,H. (2007) Statistical analysis of genotype and gene expression data. *Ph.D. Thesis*. Department of Statistics, University of Dortmund, Germany.

Schwender,H. and Ickstadt,K. (2007) Identification of SNP interactions using logic regression. *Biostatistics*, doi:10.1093/biostatistics/kxm024.

Stephens,M. and Donelly,P. (2003) A comparison of Bayesian methods for haplotype reconstruction. *Am. J. Hum. Genet*., **73**, 1162–1169.

The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.

Tusher,V. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5124.

Witte,J.S. and Fijal,B.A. (2001) Introduction: analysis of sequence data and population structure. *Genet. Epidemiol*., **21**, 600–601.