HAPLOTYPE-BASED CLASSIFIERS TO PREDICT INDIVIDUAL SUSCEPTIBILITY TO COMPLEX DISEASES

An Example for Multiple Sclerosis

María M. Abad-Grau[1], Nuria Medina-Medina[1], Andrés Masegosa[2] and Serafín Moral[2]

[1]*Departamento de Lenguajes y Sistemas Informáticos, CITIC, Universidad de Granada, Granada, Spain*

[2]*Departamento de Ciencias de la Computación e Inteligencia Artificial, CITIC, Universidad de Granada, Granada, Spain*

Keywords: Genetic predictive model, Genome-wide search, Haplotype risk.

Abstract: The enormous amount of genetic data that is currently being produced with the explosion of genome-wide association studies is yielding an important effort in the construction of genetic-based predictive models for individual susceptibility to complex diseases. However, a constant pattern of low accuracy is observed in most of them. We hypothesize that a main cause of their low accuracy is the strong reduction of genetic information considered by the classifiers, and propose a three-fold solution that considers haplotype instead of genotype individual data, whole-genome markers instead of a more stringent selection and several-marker risk variants instead of only one or two. We have compared the performance of our approach with current approaches to predict individual genetic risk to multiple sclerosis, and have found that our method yielded significantly more accurate classifiers.

1 INTRODUCTION

Building genetic-based risk models to predict individual susceptibility to a complex trait is a challenging problem that nowadays can be tackled for some complex diseases as more and more data from genome-wide association studies (GWAS) are available. However, predictive accuracy from current models seems to be very low, considering the role that genetic plays in some diseases, such as diabetes or autoimmune diseases (Wray et al., 2007; Evans et al., 2009). The very little success obtained so far may help to explain a lack in clinical application of these predictive models. Most of these genetic-based predictive model use a genetic risk score (GRS). There are two main modalities of a GRS. One is an unweighted GRS defined as the sum of all the allele risk variants ($x_i, i = 1..n$) an individual has: $GRS(x) = \sum_{i=1}^{n} x_i$, with n being the number of genetic risk positions and x_i being a three-value variable representing the genotype of an individual at position i, i.e. the number of risk variants ($0, 1$ or 2) the individual may have at this position. By considering h_i a binary variable representing the two different alleles at a given position i, $x_i = h_{i1} + h_{i2}$ holds for every $i = 1..n$, with h_{i1} and h_{i2} being respectively the two alleles making up the genotype x_i for an individual at position i. The other modality is a more accurate weighted GRS whose weights are computed as the logarithm of odds ratio at each risk position: $wGRS(x) = \sum_{i=1}^{n} w_i x_i$, with

$$w_i = \ln OR_i = \ln \frac{p(D \mid h_i = 1)}{p(\bar{D} \mid h_i = 1)} \frac{p(\bar{D} \mid h_i = 0)}{p(D \mid h_i = 0)}. \quad (1)$$

where D and $\bar{D}$ indicate an individual having or not having the disease, respectively, and h_i refers to any of the two alleles.

As an example to predict multiple sclerosis (MS) susceptibility, two different models using a weighted GRS (wGRS) have been recently published (Jager et al., 2009; Wang et al., 2011). The ability of the first of the models (Jager et al., 2009), composed of only 16 MS susceptibility loci as independent variables, to discriminate between affected and control individuals –C statistic or area under the receiver operating characteristic curve (AUC)–, was 0.64 in two different replication data sets of MS. In the second work (Wang et al., 2011), AUC rose from 0.68 to 0.769 in the replication data set when the model increased the number of independent variables from 16 to 350 genes. Still predictive capacity is too low for the model to be used for a clinical purpose. In both works, and in others performed for other complex diseases (Wray et al., 2007; Evans et al., 2009), AUC and

M. Abad-Grau M., Medina-Medina N., Masegosa A. and Moral S..
HAPLOTYPE-BASED CLASSIFIERS TO PREDICT INDIVIDUAL SUSCEPTIBILITY TO COMPLEX DISEASES - An Example for Multiple Sclerosis.
DOI: 10.5220/0003874003600366
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2012), pages 360-366
ISBN: 978-989-8425-90-4

accuracy (sensitivity and specificity) are always measured for case/control data sets.

The clinical situation highly departs from a case/control study, in which controls may have a very low number of risk loci and cases a very high number of risk loci while in the clinic, relatives of affected individuals may have a larger number of risk loci even if they are still not large enough to develop the disease. To differentiate these healthy individuals from affected ones is a much more challenging task worth for clinical purposes that is still an open research problem. Therefore in most of the cases, the population prone to need a test consists of individuals, may be newborns, with a relative having a complex disease, for whom it may be worth to know their disease susceptibility. It has to be noted, that the closer the relationship with the relative, the more specific the predictor has to be to avoid false positives. A data set that may resemble more accurately to this population is a familial trio data set with affected offspring and, usually unaffected, parents. We focused on MS, a complex disease whose genetic component is important, as there are 30% pairwise concordance for monozygotic twins (Kuusisto et al., 2008) but only 14.3% in dizygotic twins. Although there have not been many GWAS conducted on trio data sets because of the high costs required compared with case/control samples, we have been able to use genotype data from MS GWAS performed on 931 nuclear families ((IMSGC), 2010).

To build a clinically-usable predictive model for a complex disease from genome-wide data sets is a huge challenge because many loci with different relative risks may be involved in a complex disease and the environment also interacts with genetics for the final outcome. Therefore, in the construction of a predictive model, i.e., a classifier able to ascertain whether an individual will develop a complex disease, all their components must be carefully chosen.

In this paper we present results obtained when using as a risk predictor one based on the wGRS, the best of the current approaches, and show its inability to distinguish between healthy parents and affected offspring. Our conjecture to explain the lack of accuracy is that current approaches disregard genetic information by (1) only considering genotypes instead of haplotypes so that they have to use a rough inference mechanism imposed by the use of genotype data, (2) filtering loci so that only those with higher relative risk to the disease are chosen and (3) using too simple loci with only one or two simple nucleotide polymorphism (SNP) so that marker dependencies are ignored.

For those models trying the difficult task of predicting susceptibility of complex diseases using genome data, reason (1) is enough of a reason to explain their low accuracy. As a consequence of using genotypes they have to use a rough inference mechanism that yields models with a low predictive capacity. By using haplotypes, we can improve the inference mechanism so that we can assume a recessive genetic model between haplotypes, which is the genetic model on which the powerful transmission-disequilibrium test (TDT) and their multimarker extensions rely on (BickeBöller and Clerget-Darpoux, 1995; Sham and Curtis, 1995; Abad-Grau et al., 2010; Zhang et al., 2003; Yu et al., 2005; Sevon et al., 2006; Moreno-Ortega et al., 2011). In addition, by filtering loci (2) we disregard those ones with a small effect on a polygenic disease ((IMSGC), 2010) but that may be relevant for some individuals, so that sensitivity will decrease.

In the last instance, it is well-known that multimarker haplotype-based association tests (Yu et al., 2005; Abad-Grau et al., 2010) usually provide a higher power than monomarker tests as in many cases only one marker is not enough to tag a gene variant or to capture a non-recombinant variant in linkage with it. Therefore, (3) may be another reason for its low accuracy.

In this work we also develop an strategy to face these three issues. Section 2 details this strategy so that the three issues above mentioned are handled. In Section 3 we show how only by using this three-fold strategy, the predictive accuracy increases enough for the predictor to be clinically-usable. Conclusions are written in Section 4.

2 METHODS

We first describe the current state-of-the-art solutions (Wray et al., 2007; Evans et al., 2009; Jager et al., 2009) (Section 2.1). We later explain the first strategy in our solution: to use haplotypes instead of genotypes (section 2.2) and the analytical relationship between the two approaches (Section 2.3). Finally we provide a description of the way our approach goes beyond the above-mentioned simplifications (2) and (3) made by the current approaches (Section 2.4).

2.1 Currently used Predictive Models

The most widely-used approach is based on the use of genotypes and the wGRS from which a simple logistic regression model (Wang et al., 2011) is defined:

$$lnO(x) = \ln \frac{p(D \mid x)}{1 - p(D \mid x)} = \alpha_0 + \alpha_1 wGRS(x). \quad (2)$$

In terms of AUC, a classification rule based on a Naive Bayes classifier (NBC) (Domingos and Pazzani, 1997) has been shown to be equivalent to a classification rule based on a wGRS logistic regression (Equation 2) and any choice of parameters α_0 and α_1 (Sebastiani and Solovieff, 2011), i.e. this relationship is independent of the regression coefficients.

We now show that under the assumption of independent loci given the disease outcome, an assumption that yields the NBC, and by considering that $h_{ij}, j = 1,2$ are indentically distributed and are conditionally independent given D, $\alpha_1^{NBC} = 1$ and α_0^{NBC} is:

$$\alpha_0^{NBC} = \ln \frac{p(D)}{1 - p(D)} + 2 \sum_{i=1}^{n} \ln \frac{p(h_i = 0 \mid D)}{p(h_i = 0 \mid \bar{D})}, \quad (3)$$

which becomes

$$\alpha_0^{NBC} = 2 \sum_{i=1}^{n} \ln \frac{p(h_i = 0 \mid D)}{p(h_i = 0 \mid \bar{D})}, \quad (4)$$

whenever $p(D) = 1 - p(D)$.

In effect, under the assumption made by NBC, the odds of the risk of a genome-wide genotype x turns out to be:

$$O^{NBC}(x) = \frac{p(D \mid x)}{1 - p(D \mid x)} \overset{NBC}{=} \frac{p(D)}{1 - p(D)} \prod_{i=1}^{n} \frac{p(x_i \mid D)}{p(x_i \mid \bar{D})} = \frac{p(D)}{1 - p(D)} \prod_{i=1}^{n} \frac{p((h_{i1} + h_{i2}) \mid D)}{p((h_{i1} + h_{i2}) \mid \bar{D})}. \quad (5)$$

If $I_k(x_i)$ is the indicator function (i.e $I_k(x_i)$ is equal to 1/0 whether x_i is equal to k or not respectively) and by considering that h_{ij} are identically distributed and are conditionally independent given D, then the following expression holds:

$$\begin{aligned}
\ln O^{NBC}(x) &= \ln \frac{p(D)}{p(\bar{D})} + \\
&\sum_{i=1}^{n} I_0(x_i) 2 \ln \frac{p(h_i = 0 \mid D)}{p(h_i = 0 \mid \bar{D})} + \sum_{i=1}^{n} I_1(x_i) \ln \frac{p(h_i = 0 \mid D)}{p(h_i = 0 \mid \bar{D})} + \\
&\sum_{i=1}^{n} I_1(x_i) \ln \frac{p(h_i = 1 \mid D)}{p(h_i = 1 \mid \bar{D})} + \sum_{i=1}^{n} I_2(x_i) 2 \ln \frac{2p(h_i = 1 \mid D)}{2p(h_i = 1 \mid \bar{D})} \\
&= \ln \frac{p(D)}{p(\bar{D})} + 2 \sum_{i=1}^{n} \ln \frac{p(h_i = 0 \mid D)}{p(h_i = 0 \mid \bar{D})} \\
&- \sum_{i=1}^{n} x_i \ln \frac{p(h_i = 0 \mid D)}{p(h_i = 0 \mid \bar{D})} + \sum_{i=1}^{n} x_i \ln \frac{p(h_i = 1 \mid D)}{p(h_i = 1 \mid \bar{D})} \\
&= \ln \frac{p(D)}{p(\bar{D})} + 2 \sum_{i=1}^{n} \ln \frac{p(h_i = 0 \mid D)}{p(h_i = 0 \mid \bar{D})} \\
&+ \sum_{i=1}^{n} x_i \ln \frac{p(\bar{D} \mid h_i = 0)}{p(D \mid h_i = 0)} \frac{p(D \mid h_i = 1)}{p(\bar{D} \mid h_i = 1)}, \quad (6)
\end{aligned}$$

being the first two addends the intercept α_0^{NBC} and the last one the weighted genetic risk score $wGRS(x)$, so that the coefficient $\alpha_1^{NBC} = 1$.

A simpler genotype-based approach assumes $\alpha_1 = 1$ and $\alpha_0 = 0$ (Jager et al., 2009).

We compared both models with our *haplotype-based absolute-risk recessive model* approach explained below. To avoid zero probability values because of small sample sizes, we estimated probabilities $p(D)$ and $p(h_i), i = 1..n$ by using a Bayesian estimator, and considered a discrete uniform distribution with $n = 1$ as the prior distribution for all of them, in all the approaches.

2.2 Strategy 1: Haplotype-based Absolute-risk Recessive Model

We performed two modifications to the simple logistic regression model: a haplotype-based approach instead of a genotype-based approach and a recessive model on the absolute risk of the genome-wide haplotypes instead of a multiplicative model of the genome-wide haplotypes on the odds of the disease.

2.2.1 Haplotype-based Approach

We will first introduce the concept of genetic risk score of an haplotype (hwGRS), i.e. the relative risk score of the genetic material the individual inherits from only one of their two parents:

$$hwGRS(h_j) = \sum_{i=1}^{n} w_i h_{ij}, j = 1,2. \quad (7)$$

The relationship between the two haplotype-based scores and the genetic score of an individual is:

$$\begin{aligned}
wGRS(x) &= hwGRS(h_1) + hwGRS(h_2) \\
&= \sum_{i=1}^{n} w_i (h_{i1} + h_{i2}) = \sum_{i=1}^{n} w_i h_{i1} + w_i h_{i2}, \quad (8)
\end{aligned}$$

with h_{i1} and h_{i2} being the two haplotypes making up the individual's genotype x_i.

The odds for the disease are computed independently for each of the two genome-wide haplotypes, h_1, h_2, the genome-wide genotype x of an individual has:

$$lnO(h_j) = \ln \frac{p(D \mid h_j)}{1 - p(D \mid h_j)} = \alpha_0^h + \alpha_1^h hwGRS(h_j), \quad (9)$$

$j = 1,2$.

As it has been done for the genotype-based approach, the intercept α_0^h and α_1^h were computed by assuming loci are independent given the disease outcome (i.e. by using NBC). Under this assumption, and in the case of window size of 1, it is straightforward to show that $\alpha_1^{NBCh} = 1$ and

$$\alpha_0^{NBCh} = \ln \frac{p(D)}{1-p(D)} + \sum_{i=1}^{n} \ln \frac{p(h_i = 0 \mid D)}{p(h_i = 0 \mid D)}, \quad (10)$$

which becomes

$$\alpha_0^{NBCh} = \sum_{i=1}^{n} \ln \frac{p(h_i = 0 \mid D)}{p(h_i = 0 \mid D)}, \quad (11)$$

whenever $p(D) = 1 - p(D)$.

The relationship between the haplotype odds and the genotype odds for the disease is:

$$\ln O^{NBC}(x) = \ln O^{NBCh}(h_1) + \ln O^{NBCh}(h_2) - \ln \frac{p(D)}{1-p(D)}, \quad (12)$$

and it becomes

$$\ln O^{NBC}(x) = \ln O^{NBCh}(h_1) + \ln O^{NBCh}(h_2) \quad (13)$$

whenever $p(D) = 1 - p(D)$.

2.2.2 Absolute-risk Models

However, our proposal also changes the inference procedure, so that the odds of the disease an individual has is not computed as the product of the odds for each haplotype, as it occurs in the commonly used approaches. In those approaches, the absolute individual risk is computed by assuming a multiplicative effect on the odds of each haplotype (see Equation 12):

$$\frac{p(D \mid h_1)}{1-p(D \mid h_1)} \frac{p(D \mid h_2)}{1-p(D \mid h_2)} \frac{p(D)}{1-p(D)}. \quad (14)$$

In the case $p(D) = 1 - p(D)$ holds it will become:

$$O(X) = \frac{p(D \mid x)}{1-p(D \mid x)} = O(h_1)O(h_2) =$$

$$\frac{p(D \mid h_1)}{1-p(D \mid h_1)} \frac{p(D \mid h_2)}{1-p(D \mid h_2)}. \quad (15)$$

The main problem of this approach is that the odds of the disease for an individual with genotype x only considers the odds of having the two genome-wide haplotypes being high-risk haplotypes versus being low-risk haplotypes but it is disregarding those cases in which the genotype may have one high risk haplotype and other low risk haplotype.

Instead of that, we apply the genetic model on the absolute individual risks, as we are assessing the absolute individual risk in order to infer individual disease susceptibility. Depending on the genetic model assumed between haplotypes (recessive, additive and dominant) we have defined three different modalities explained below.

Absolute-risk Recessive Model. Therefore, assuming a recessive effect on the haplotype risks, i.e. the same assumption done by the TDT so that the two genome-wide haplotypes have to be considered of high risk for the individual to be at risk:

$$p(D \mid x) = p(D \mid h_1)p(D \mid h_2) \quad (16)$$

so that we compute:

$$\frac{p(D \mid x)}{1-p(D \mid x)} = \frac{p(D \mid h_1)p(D \mid h_2)}{1-p(D \mid h_1)p(D \mid h_2).} \quad (17)$$

Absolute-risk Additive Model. Other genetic model assumptions is the additive model, which means

$$\frac{p(D \mid x)}{1-p(D \mid x)} = \frac{p(D \mid h1) + p(D \mid h2)}{1-p(D \mid h1) - p(D \mid h2)}. \quad (18)$$

Absolute-risk Dominant Model. Under the dominant model, in which at least one high risk haplotype is required to have the disease,

$$p(D \mid x) = p(D \mid h_1)p(D \mid h_2) \begin{array}{r} +p(\bar{D} \mid h_1)p(\underline{D} \mid h_2) \\ +p(D \mid h_1)p(\bar{D} \mid h_2) \end{array} \quad (19)$$

and thus $\frac{p(D|x)}{1-p(D|x)}$ is computed as

$$\frac{p(D \mid h_1)p(D \mid h_2) \begin{array}{r} +p(\bar{D} \mid h_1)p(\underline{D} \mid h_2) \\ +p(D \mid h_1)p(\bar{D} \mid h_2) \end{array}}{p(\bar{D} \mid h_1)p(\bar{D} \mid h_2)}. \quad (20)$$

In our experiments, we only used the recessive model, as it is the model on which TDT relies on and the test has been proved to be very powerful to detect association loci in several GWAS conducted on polygenic diseases, including MS.

2.3 Analytical Relationship between Genotype-based Classifiers and Haplotype-based Classifiers

In order to reduce computation time, we have analytically assessed the relationship between the estimation of individual risk made by the state-of-the-art genotype-based classifier ($p_G(D \mid x)$) and by our haplotype-based classifier ($p_H(D \mid x)$), under the

assumption of conditional independence made by NBC. Therefore, we will compute both of them from the haplotype-based weighted GRS (hwGRS) defined above.

2.3.1 Genotype-based Predictors built from the Haplotype-based and Weighted hwGRS

By considering Equation 12, it is straightforward to show that

$$p_G(D \mid x) = \frac{1}{1+e^{-\ln O^{NBC}(x)}} = \frac{1}{1+e^{-\alpha_0^{NBC}-wGRS(x)}}$$
$$= \frac{1}{1+e^{-\alpha_0^{NBC}-hwGRS(h_1)-hwGRS(h_2)}}$$
$$= \frac{1}{1+e^{-\ln O^{NBC}(h_1)-\ln O^{NBC}(h_2)+\ln\frac{p(D)}{p(D)}}}, \quad (21)$$

with

$$\alpha_0^{NBC} = 2\sum_{i=1}^{n} \ln \frac{p(x_i=0 \mid D)}{p(x_i=0 \mid D)} + \ln \frac{p(D)}{1-p(D)}.$$

In the situation $p(D) = 1-p(D)$, it becomes:

$$\frac{1}{1+e^{-\ln O^{NBC}(h_1)-\ln O^{NBC}(h_2)}}. \quad (22)$$

In the situation of no intercept, the risk is estimated as:

$$p_G(D \mid x) = \frac{1}{1+e^{-\ln O^{NBC}(x)}} = \frac{1}{1+e^{-wGRS(x)}}$$
$$= \frac{1}{1+e^{-hwGRS(h_1)-hwGRS(h_2)}}. \quad (23)$$

2.3.2 The Haplotype-based Predictor built from the Haplotype-based and Weighted hwGRS

As explained above, the individual predictive model was defined by assuming a recessive genetic model, i.e. a multiplicative effect of the haplotype risks (Equation 17) and therefore by combining them to obtain the individual risk. An individual-risk classifier has a binary class representing whether the individual has the disease. The final score $p(D \mid x)$ represents the probability for an individual susceptibility to MS and is obtained by computing the joint probability for an individual of having both high risk haplotypes. We computed the probability of having the two high risk haplotypes because TDT assumes that both haplotypes in affected individuals are high risk haplotypes while in unaffected parents, only one haplotype is a high risk haplotype.

$$p_H(D \mid x) = p(D \mid h_1)p(D \mid h_2)$$
$$= \left(\frac{1}{1+e^{-\alpha_0^{NBCh}-hwGRS(h_1)}}\right)\left(\frac{1}{1+e^{-\alpha_0^{NBCh}-hwGRS(h_2)}}\right)$$
$$= \left(\frac{1}{1+e^{-\ln O^{NBC}(h_1)}}\right)\left(\frac{1}{1+e^{-\ln O^{NBC}(h_2)}}\right) \quad (24)$$

with

$$\alpha_0^{NBCh} = \sum_{i=1}^{n} \ln \frac{p(h_i=0 \mid D)}{p(h_i=0 \mid D)} + \ln \frac{p(D)}{1-p(D)}. \quad (25)$$

2.4 Strategies 2 and 3: Multimarker Variables and Loci Selection

Instead of considering single markers, we tested the three approaches by grouping consecutive markers into binary variables (low and high risk) so that each variant is coded as high or low risk by using the 2G algorithm (Abad-Grau et al., 2011). 12 different amounts of consecutive markers were tried: $1,2,5,10,15,20,25,30,40,50,100,150$. To test the effect of loci selection (strategy 3), we performed different levels of loci filtering, by imposing different upper-limits in the p-value (0.8, 0.6, 0.4, 0.2, 0.15, 0.1, 0.050, 1e-2, 1e-3, 1e-4, 1e-5, 1e-6, 1e-7) obtained by the multimarker TDT $mTDT_{2G}$ (Abad-Grau et al., 2011), which measures whether there are significant differences in the amount of high risk haplotypes that are transmitted to the affected offspring compared with those non-transmitted ones.

Therefore TDT_{2G} was applied along the genome by using sliding windows with the 12 different configurations of window sizes and offset of 1 and the 13 p value upper limits above mentioned. By combining the 12 different window sizes, the 13 different p value upper limits and the three approaches compared we produced $12 \times 13 \times 3$ predictive models. Since our approach was based on haplotypes and the relationship between haplotype-based and genotype-based odds in the current approaches has been established above, we estimated as a first step the log odds for each genome-wide haplotype. We considered as high/low risk genome-wide haplotypes those that were transmitted/not transmitted to the offspring (the 22 autosome transmitted/non-transmitted chromosomes).

In a second step, the log odds for each genome-wide homologous chromosome of an individual were combined in order to estimate the individual risk $p(D \mid x)$ by considering

$$p(D \mid x) = \frac{1}{1+e^{-\ln O(x)}} \quad (26)$$

holds in a logistic regression model, being $O(x)$ the odds of $D \mid x$, and the way each of the methods work.

Once individual predictive models were built, we used them to measure their generalization capacity,

i.e., how accurate would be a model when used in a different data set. To encode, in a replication data set, whether a haplotype at a given sliding window was a high (1) or low (0) risk one, we computed the similarity between it and every haplotype in the list of high risk and low risk haplotypes for the corresponding sliding window in the training data set. Therefore, we classified it as 1 or 0 depending on whether the closest haplotype belonged to the set of high or low risk haplotypes respectively. For the similarity measure we used the length measure (Tzeng et al., 2003), which computes the largest number of consecutive matching alleles.

3 RESULTS

We computed the training and predictive accuracies and C-statistics (AUC) for all the individual-risk predictive models built using the 13 different p-value upper limits and 12 different window sizes. Predictive accuracies and AUCs are results obtained when different independent data sets are used to learn the model and to predict risk. We randomly selected 500 family trios as the training data set and the remaining 431 as the test data set.

Results (predictive accuracy) shown in Figure 1 compares results between our approach (values at the right side of the "Haplotype-all" and brown line) and the standard multiplicative genotype-based model using wGRSs ("Genotype-filtering" and blue line). The current approach is not able to perform a good prediction and neither did it succeed when we applied only the strategy of no genetic filtering ("Genotype-all" and orange line). However, we found a substantial accuracy increase when trying the three-fold strategy at a time consisting on a recessive haplotype-based model with no filtering and large multimarker variables (values at the right side of the "Haplotype-all" and brown line) instead of shorter multimarker variables or single marker variables (values at the left side), or risk-based filtering ("Haplotype-filtering" and green line) alone. We used NBC to build all the models. To make sure that the haplotype-based strategy was crucial for the results obtained, we randomly flipped positions and found an AUC decay to values around 0.5 (data not shown). Figure 2 compares results between our approach and the standard multiplicative genotype-based model using wGRSs and no intercept. With this plot we wanted to check whether the intercept is important for the outcome and, as it can be seen in the plot, accuracy is lower than when using the intercept (Figure 1).

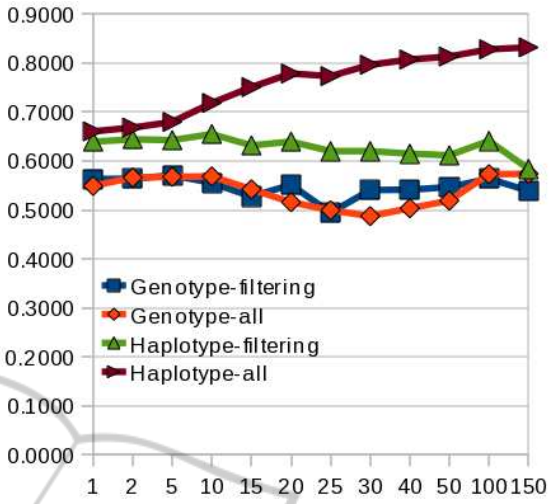

Figure 1: Accuracy (y-axis) of genetic predictors for different sizes of multimarker variables (x-axis). Comparison between our approach and other logistic regression models using the intercept.

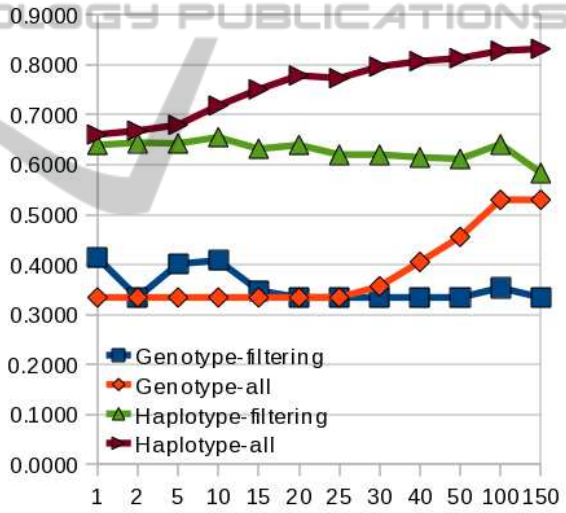

Figure 2: Accuracy (y-axis) of genetic predictors for different sizes of multimarker variables (x-axis). Comparison between our approach and other logistic regression models disregarding the intercept.

4 CONCLUSIONS

These results shed light on the challenging task of building predictive models of individual genetic risk in complex diseases. Using allelic association between SNPs, a recessive genetic model and several markers at a time seem to be all essential to obtain a disease predictor enough accurate to be used in the clinic. However, this is only a first step in a new direction in the search of genome-wide predictors of

complex diseases. New wet-lab or in-silico methods to accurately reconstruct very long haplotypes instead of using the expensive nuclear family data sets have to be defined. Moreover, the method needs to be tested in other polygenic diseases.

ACKNOWLEDGEMENTS

The authors were supported by the Spanish Research Program under project TIN2010-20900-C04, the Andalusian Research Program under project P08-TIC-03717 and the European Regional Development Fund (ERDF). The authors thank Paola Sebastiani for her help in the work undertaken.

REFERENCES

Abad-Grau, M., Medina-Medina, N., Montes-Soldado, R., Matesanz, F., and Bafna, V. (2011). Sample reproducibility of genetic association using different multimarker tdts in genome-wide association studies: Characterization and a new approach. *PLoS ONE*, accepted.

Abad-Grau, M., Medina-Medina, N., Montes-Soldado, R., Moreno-Ortega, J., and Matesanz, F. (2010). Genome-wide association filtering using a highly locus-specific transmission/disequilibrium test. *Human Genetics*, 128:325–44.

BickeBöller, H. and Clerget-Darpoux, F. (1995). Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. *Genet Epidemiol*, 12:865–70.

Domingos, P. and Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29:103–37.

Evans, D., Visscher, P., and Wray, N. (2009). Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Human Molecular Genetics*, 18:3525–31.

(IMSGC), I. M. S. G. C. (2010). Evidence for polygenic susceptibility to multiple sclerosis - the shape of things to come. *Am J Hum Genet*, 86:621–5.

Jager, P. D., Chibnik, L., Cui, J., Reischl, J., Lehr, S., Simon, K., Aubin, C., Bauer, D., Heubach, J., Sandbrink, R., Tyblova, M., Lelkova, P., 'Steering committee of the BENEFIT study, committee of the BEYOND study', S., committee of the LTF study', S., committee of the CCR1 study', S., E, E. H., Pohl, C., Horakova, D., Ascherio, A., Hafler, D., and Karlson., E. (2009). Integration of genetic risk factors into a clinical algorithm for multiple sclerosis susceptibility: a weighted genetic risk score. *Lancet Neurol.*, 8(12):1111–9.

Kuusisto, H., Kaprio, J., Kinnunen, E., Luukkaala, T., Koskenvuo, M., and Elovaara, I. (2008). Concordance and heritability of multiple sclerosis in finland: study on a nationwide series of twins. *Eur J Neurol.*, 15(10):1106–10.

Moreno-Ortega, J. J., Medina-Medina, N., Montes-Soldado, R., and Abad-Grau, M. M. (2011). Improving reproducibility on tree based multimarker methods: Treedth. In Rocha, M., Corchado, J., Fernández-Riverola, F., and Valencia, A., editors, *PACBB '11: Proceedings of the 5th International Conference on Practical APplications of Computational Biology and Bioinformatics*, volume 1, pages 1–8, Berlin, Heidelberg. Springer-Verlag.

Sebastiani, P. and Solovieff, N. (2011). Nave bayesian classifier and genetic risk score for genetic risk prediction of a categorical trait: Not so different after all! *submitted*.

Sevon, P., Toivonen, H., and Ollikainen, V. (2006). Treedt: Tree pattern mining for gene mapping. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 3(2):174–85.

Sham, P. C. and Curtis, D. (1995). An extended transmission/disequilibrium test (tdt) for multiallelic marker loci. *Annals of Human Genetics*, 59:323–336.

Tzeng, J., Devlin, B., Wasserman, L., and Roeder, K. (2003). On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am J Hum Genet*, 72:891–902.

Wang, J. H., Pappas, D., Jager, P. L. D., Pelletier, D., de Bakker, P. I., Kappos, L., Polman, C. H., 'Australian, (ANZgene)', N. Z. M. S. G. C., Chibnik, L. B., Hafler, D. A., Matthews, P. M., Hauser, S. L., Baranzini, S. E., and Oksenberg, J. R. (2011). Modeling the cumulative genetic risk for multiple sclerosis from genome-wide association data. *Genome Medicine*, 3:3.

Wray, N., Goddard, M., and Visscher, P. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research*, 17:1520–28.

Yu, K., Gu, C. C., Xiong, C., An, P., and Province, M. (2005). Global Transmission/Disequilibrium tests based on haplotype sharing in multiple candidate genes. *Genetic Epidemiology*, 29:223–35.

Zhang, S., Sha, Q., Chen, H., Dong, J., and Jiang, R. (2003). Transmission/Disequilibrium test based on haplotype sharing for tightly linked markers. *Am J Hum Genet*, 73:566–79.