Riskoweb: web-based genetic profiling to complex disease using genome-wide SNP markers

Sergio Torres-Sánchez, Rosana Montes-Soldado, Nuria Medina-Medina, Andrés R. Masegosa and María Mar Abad-Grau

Abstract Assessing risk susceptibility of individuals to a complex disease is becoming an interesting prevention tool specially recommended for those with ancestors or other relatives affected by the disease. As genome-wide DNA sequencing is getting more affordable, more dense genotyping is performed and accuracy is increased. Therefore, health public services may consider the results of this approach in their preventing plans and physicians be encouraged to perform these risk tests. A web-based tool has been built for risk assessing of complex diseases and its knowledge base is currently filled with multiple sclerosis risk variants and their effect on the disease. The genetic profiling is calculated by using a Naive Bayes network, which it has been shown to provide highly accurate results as long as dense genotyping, haplotype reconstruction and several markers at a time are considered.

Sergio Torres-Sánchez
Department of Computer Languages and Systems - CITIC - University of Granada. e-mail: s.torres.sanchez@gmail.com

Rosana Montes-Soldado
Department of Computer Languages and Systems - CITIC -University of Granada. e-mail: rosana@ugr.es

Nuria Medina-Medina
Department of Computer Languages and Systems - CITIC -University of Granada. e-mail: nmedina@ugr.es

Andrés R. Masegosa
Department of Computer Science and Artificial Intelligence - CITIC -University of Granada. e-mail: andrew@decsai.ugr.es

María Mar Abad-Grau
Department of Computer Languages and Systems - CITIC -University of Granada. e-mail: mabad@ugr.es

1 Introduction

Although it does not exist yet an algorithm to assess the individual genetic risk to most complex diseases or traits, several genome-wide association studies (GWASs) for different traits are currently being performed in different labs all over the world. Moreover, the number of GWASs being tackled is increasing as genetic sequenciation becomes more affordable.

At the moment, genetic predictors of individual susceptibility have been built for research purpose for a few complex diseases with a large genetic component, such as type I diabetes or rheumatoid arthritis [10, 1] reaching a predictive accuracy between 65% and 75%. However, their use for any complex disease or trait is not yet being considered as a practical choice for physicians or individuals with affected relatives that may be interested in being tested for disease susceptibility. Two main reasons may be on the basis of this issue: the high cost of sequencing an individual genome and the low accuracy of most predictors that makes them to still being under development phase. However, with the cost decrease in genome sequencing more GWASs are being performed and high accuracies may be reached for complex diseases with lower genetic component. Still, once a predictor is built, and as new GWASs may improve predicting accuracy, the knowledge base is constantly updated and this may make highly difficult to think about distributing predictors as stand alone software.

In this work we have built a novel web-based predictor whose knowledge base has been currently filled with genetic variants to Multiple Sclerosis (MS) able to assess individual susceptibility to MS with a 81.69% accuracy. This high accuracy has been reached using thousand of risk variants composed by more than 20 markers each one. Moreover, accuracy is expected to improve with more dense GWASs, larger samples and more markers in the risk variants are used. Because of that, the web application has been developed as an evolutive tool so that the knowledge base can be updated by experts with information coming from more dense and or larger sample size GWASs. It also allows to introduce genetic variants for any other complex disease and to automatically build a predictive model (the knowledge base) based on a Naive Bayes classifier.

We explain in Section 2 the data source and the model used to store knowledge about disease susceptibility. Section 3 is devoted to describe the functionality of the web-based application. Conclusions are explained in Section 4.

2 Method

To build the predictive model we focused on MS, as we had access to the raw data from a genome-wide association study performed by the International

Multiple Sclerosis Genetic Consortium using a DNA microarray –GeneChip Human Mapping 500K ArraySet by Affymetrix– to examine 334,923 single nucleotide polimorphism (SNP) markers [4] in 931 family trios. Using family trios instead of unrelated individuals arises accuracy in the process of haplotype reconstruction from genotypes, a very important step under our approach, as other approaches which do only consider genotypes reported lower accuracies [1, 5]. Once haplotypes were reconstructed from genotypes by using family information and the expectation-maximization (EM) algorithm in case of ambiguity [11], a multimarker transmission-disequilibrium test (TDT) which groups haplotypes by low and high risk [7], was genome-wide applied by using different sizes of overlapping windows of SNPs (sliding windows) and offset of 1. We found that accuracy improved with an increase in window size and an increase on the p-value upper bound (i.e., by relaxing the criterion to consider a window as a risk locus), in agreement with a recent work which considers MS as a complex disease with thousand small variants and thousand very small effects along the genome [3]. Figure 2 shows the classifier performance under different p-value upper limits (x-axis) and window sizes (y-axis). We built the predictor using sliding windows of size 20 and p value upper limit of 0.001 because this configuration reached a performance (measured by the area under the receptor-operative-curve (AUC) or C-statistic) of 81.69%. The prediction is performed in two phases 1: First a genome-wide haplotype predictor of disease susceptibility is computed for each of the two haplotypes of an individual. Second, the individual predictor of disease susceptibility multiplies the two outputs of the haplotype predictor (see Figure 1).

To represent the predictive model, we tried several approaches such as Bayesian networks, instance-based measures, support vector machines (SVM) [9], decision trees (DT) and random forests (RF). Among the different Bayesian network-based algorithms to build classifiers that we tried, the simplest one, Naive Bayes (NB), was the only one computationally affordable and the one which achieved the highest accuracy. Among the other approaches we disregarded instance-based measures for being highly time consuming for this particular problem, which requires thousand of variables (risk loci) to achieve a high performance. Figure 3 shows the results (AUC and accuracy) returned by the haplotype risk predictor using some computationally affordable state-of-art classification algorithms using the parameters with the highest performance: NB, a SVM algorithm with a sigmoid kernel function, RF with 4000 trees, a boosting algorithm (AdaBoostM1) [2] with 20000 trees and c5.4 [6]. Therefore, we decided to use NB to build the predictive model, as the model reached the highest performance, is simple and easy to interpret.

3 Functionality

Riskoweb is a web-based evolutive application with the following functions: (1) builds a predictive model, which is a Naive Bayes classifier, for a new disease or a new population from a data set with SNP markers of nuclear families, (2) updates an existing predictive model with new risk loci, (3) computes the individual genetic profiling to a complex disease and (4) graphically displays an individual risk map for all the risk loci used by the model, a plot which has been called a riskomap. The application evolves by updating the knowledge base as new data is introduced. Therefore, discoveries of new risk loci to a disease may be introduced so that the predictive overall accuracy will improve.

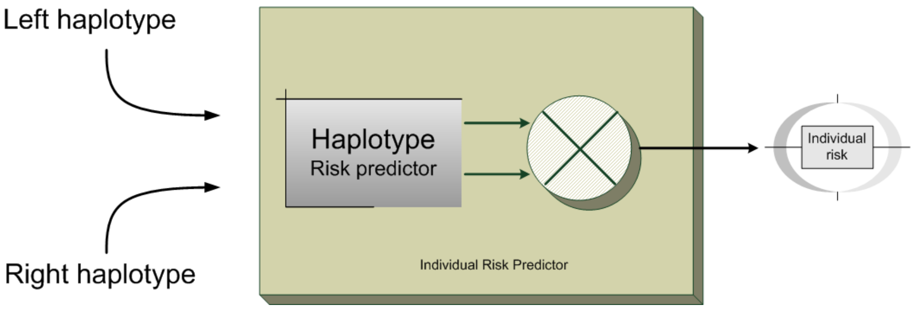


Fig. 1 Arquitecture of the individual risk predictor used to assess the individual susceptibility to complex diseases.

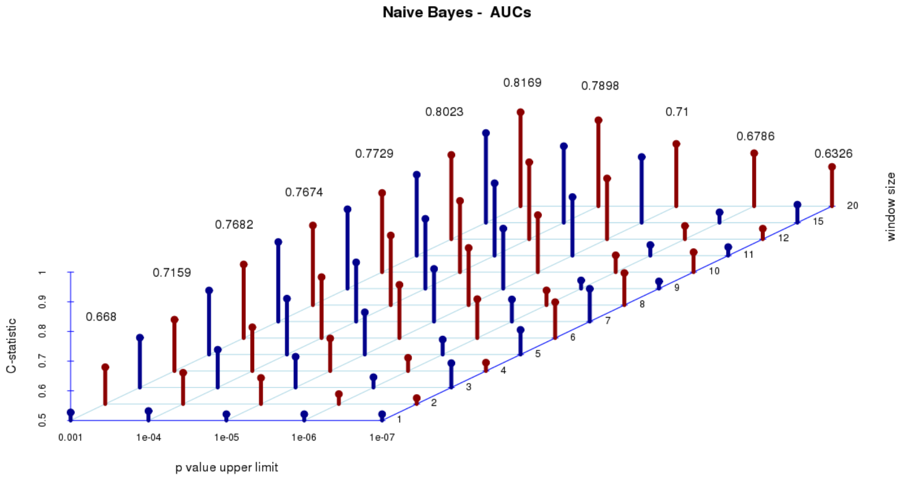


Fig. 2 C statistic (AUC) reached by the individual risk predictor using the NB algorithm with different p value upper bounds (x-axis) and window sizes (y-axis).

1. Building a predictive model: To build a predictive model for a new complex disease or another population of a disease already existing in the knowledge base, two sets of files have to be provided. One set with information about the SNP genotypes and pedigree information used to learn the parameters for the model and the other set with information with the risk loci that have been found in association with the disease and will therefore be used to build the model structure. The first set of files consists of three files for each somatic chromosome: one in extended makeped format, a widely used format to store individual phenotypes and genotypes, with phenotype and genotype information for a set of nuclear families (parents and an affected offspring) and the other two to feed the knowledge base with the physical and rs number (a unique SNP identification regardless the SNP assembly used for individual genotype sequencing) positions of the SNPs used in the first file. Files in the second set, containing information about every risk locus has also to be provided. For each risk locus another three files are required: one with the rs numbers and chromosome, and the other two with a list of high/low risk haplotypes at that locus, which usually consists of a few consecutive SNPs.
2. Updating a predictive model with a new risk locus: New loci affecting a disease onset may be discovered. The knowledge base will be updated when information about the locus is introduced by using the three files described above for a risk locus. Once those files are introduced, Riskoweb will perform the following tasks: (2.1) it will first extract for each chromosome and each individual the pair of haplotypes from each genotype. Family-based information will be used for haplotype reconstruction and, in case of ambiguity, the EM algorithm [11], (2.2) it will compute for each new risk locus a list of high and low risk haplotypes by using a TDT algorithm to compare differences in transmission counts, i.e., whether an

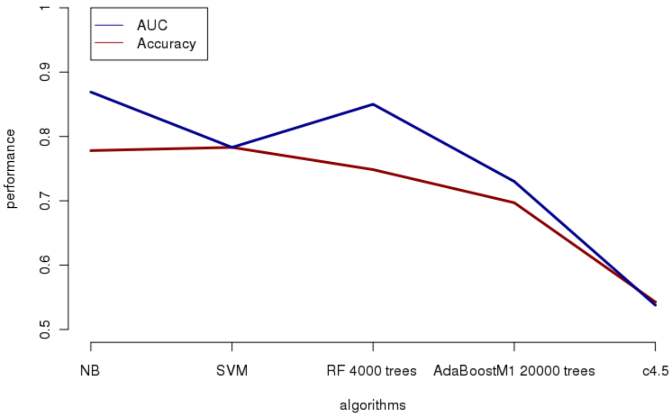


Fig. 3 Accuracies (red line) and AUCs (blue line) reached by the haplotype risk predictor using different learning machines for p value=0.001 and windoew size 15.

haplotype from a parent is more often transmitted than non-transmitted, (2.3) it will code haplotypes using a binary code according to whether they are considered high or low risk and (2.4) it will compute the parameters for the NB classifier using haplotype counts in the data set and whether the haplotype is transmitted or not (the class variable for the haplotype risk predictor).

3. Computing individual risk: When a genome-wide extended-makeped file for only one individual is provided together with a file containing information about the rs number of each genotype, the application performs the following tasks: (3.1) it first extracts the pair of genome-wide haplotypes for this individual from their genotype, (3.2) it extracts from each genome-wide haplotype those SNPs considered at risk positions by the predictive model, (3.3) it then computes for each risk position and each haplotype whether its a high-risk or a low-risk haplotype, depending on its similarity to high and low risk haplotypes in the model (to test haplotype similarity, it uses the length measure [8], which computes the largest number of consecutive matching alleles) and (3.4) it introduces the two final binary-coded haplotypes in the individual risk predictor (see Figure 1) and returns the probability for this individual to develop the complex disease being tested.
4. Plotting a riskomap: A riskomap has 22 columns (one for each somatic chromosome), and the height of the columns is proportional to the number of risk loci used by the predictive model. The image of a riskomap is formed by green, blue and red cells, meaning homozygous for low-risk haplotype, heterozygous and homozygous for high-risk haplotype, respectively (see Figure 4 as an example).

There are three types of users in the web application: registered user, privileged user and admin. Without registration, a casual user just can read the information displayed on the site (see Figure 5), but they will not be able to interact with it in any way. As several specific data is required for

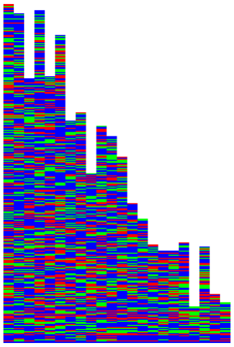
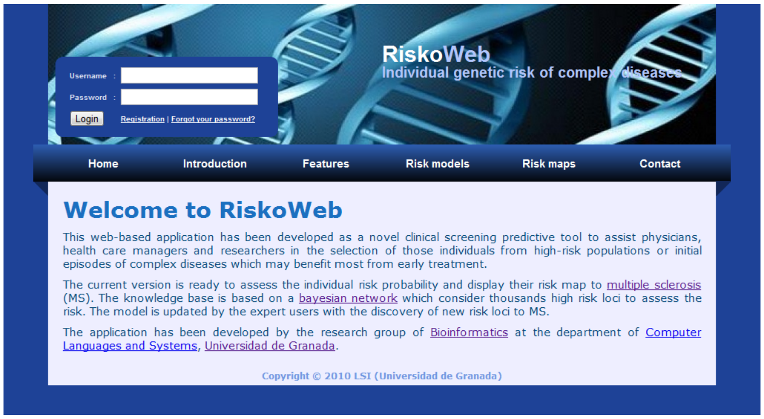


Fig. 4 An example of an individual risk map or riskomap to MS.

registration, we can check who is using the system and for what institution are they working (universities, research groups, etc.). A registered user has the permission to upload the genetic information of a person, and the system will create a riskomap with the data. Furthermore, any registered user has the option to ask for a promotion to privileged user. Privileged users can generate riskomaps too, and in addition they can create new risk models and modify the risk models that already exist. Lastly, an admin is just like a privileged user, except that when a registered user ask for a promotion, only an admin can accept or deny that request.

4 Conclusions

We have built a web-based application to assess individual susceptibility to a complex disease. The predictor combines results of genome-wide haplotype risks computed by using a Naive Bayes classifier and it returns the individual risk, a probability, and a graphical representation of the individual risk for all the loci considered by the model, the riskomap. Its knowledge base is currently equipped with a model to predict individual susceptibility to MS. This web resource can be easily used by physicians and researchers and has evolutive capabilities so that it can tackle with the speed to which new genetic variants are being discovered. This novel predictive tool to perform clinical screening may assist physicians, health care managers and researchers in the selection of those individuals from high-risk populations or with initial episodes of a complex disease who may benefit most from early treatment.



Welcome to RiskoWeb

This web-based application has been developed as a novel clinical screening predictive tool to assist physicians, health care managers and researchers in the selection of those individuals from high-risk populations or initial episodes of complex diseases which may benefit most from early treatment.

The current version is ready to assess the individual risk probability and display their risk map to multiple sclerosis (MS). The knowledge base is based on a bayesian network which consider thousands high risk loci to assess the risk. The model is updated by the expert users with the discovery of new risk loci to MS.

The application has been developed by the research group of Bioinformatics at the department of Computer Languages and Systems, Universidad de Granada.

Copyright © 2010 LSI (Universidad de Granada)

Fig. 5 Homepage of riskoweb

Web resources

The website has been created at http://bios.ugr.es/riskoweb.

Acknowledgements The authors were supported by the Spanish Research Program under project TIN2007-67418-C03-03, the Andalusian Research Program under project P08-TIC-03717 and the European Regional Development Fund (ERDF).

References

1. Evans, D., Visscher, P., Wray, N.: Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Human Molecular Genetics **18**, 3525–31 (2009)
2. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference on Machine Learning, pp. 148–56 (1996)
3. (IMSGC), I.M.S.G.C.: Evidence for polygenic susceptibility to multiple sclerosis - the shape of things to come. Am J Hum Genet **86**, 621–5 (2010)
4. 'International Multiple Sclerosis Genetics Consortium', D.H., Compston, A., Sawcerand, S., Lander, E., Daly, M., Jager, P.D., de Bakker, P., Gabriel, S., Mirel, D., Ivinsonand, A., Pericak-Vance, M., Gregory, S., Rioux, J., McCauley, J., Haines, J., Barcellos, L., Cree, B., Oksenberg, J., Hauser, S.: Risk alleles for multiple sclerosis identified by a genomewide study. New England Journal of Medicine **357**(9), 851–62 (2007)
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., Karlson., E.: Integration of genetic risk factors into a clinical algorithm for multiple sclerosis susceptibility: a weighted genetic risk score. Lancet Neurol. **8**(12), 1111–9 (2009)
6. Quinlan, R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc. (1993)
7. Schaid, D.: General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol **1996**, 423–449 (1996)
8. TZeng, J., Devlin, B., Wasserman, L., Roeder, K.: On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am J Hum Genet **72**, 891–902 (2003)
9. Vapnik, V.: The Nature of Statistical Learning Theory. Springer (1999)
10. Wray, N., Goddard, M., Visscher, P.: Prediction of individual genetic risk to disease from genome-wide association studies. Genome Research **17**, 1520–28 (2003)
11. Zhang, S., Sha, Q., Chen, H., Dong, J., Jiang, R.: Transmission/Disequilibrium test based on haplotype sharing for tightly linked markers. Am J Hum Genet **73**, 566–79 (2003)