BIOINFORMATICS APPLICATIONS NOTE

Vol. 25 no. 9 2009, pages 1185–1186
doi:10.1093/bioinformatics/btp121

Genome analysis

A web server for inferring the human *N*-acetyltransferase-2 (NAT2) enzymatic phenotype from *NAT2* genotype

Igor B. Kuznetsov*, Michael McDuffie and Roxana Moslehi*

Gen*NY*Sis Center for Excellence in Cancer Genomics, Department of Epidemiology and Biostatistics,
University at Albany, One Discovery Drive, Rensselaer, NY 12144, USA

Received on November 26, 2008; revised on February 5, 2009; accepted on February 26, 2009

Advance Access publication March 4, 2009

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: *N*-acetyltransferase-2 (NAT2) is an important enzyme that catalyzes the acetylation of aromatic and heterocyclic amine carcinogens. Individuals in human populations are divided into three NAT2 acetylator phenotypes: slow, rapid and intermediate. NAT2PRED is a web server that implements a supervised pattern recognition method to infer NAT2 phenotype from SNPs found in *NAT2* gene positions 282, 341, 481, 590, 803 and 857. The web server can be used for a fast determination of NAT2 phenotypes in genetic screens.

Availability: Freely available at http://nat2pred.rit.albany.edu

Contact: ikuznetsov@albany.edu; rmoslehi@albany.edu

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

1 INTRODUCTION

N-acetyltransferase-2 (NAT2) is an important enzyme that catalyzes the acetylation of aromatic and heterocyclic amine carcinogens (Blum *et al.*, 1990). Based on the level of NAT2 acetylator activity, individuals in human populations are divided into three enzymatic phenotypes: rapid (normal activity), intermediate and slow (reduced activity) (Hein *et al.*, 2000). Single nucleotide polymorphisms (SNPs) within *NAT2* determine the NAT2 acetylator phenotype. A consensus has been reached on association between *NAT2* genotype and acetylator phenotype (Hein, 2006). Recently, we showed that individuals with *NAT2* SNP variants associated with the slow phenotype were more susceptible to the effects of tobacco smoking with respect to the risk of developing an advanced colorectal adenoma (Moslehi *et al.*, 2006). Several other studies have also linked *NAT2* gene variants and acetylator phenotypes to the risk of several malignant and pre-malignant conditions (Brockton *et al.*, 2000; Hein, 2006; Potter *et al.*, 1999; Tiemersma *et al.*, 2004). The identification of at-risk individuals is an important component of cancer prevention. Current genotyping technologies are able to determine which alleles are present at each locus, but do not provide information about the phase of the alleles at different loci (i.e. do not provide information about which alleles at adjacent loci occur on the same chromosome). In order to assign an acetylator phenotype to a particular individual, the *NAT2* haplotypes for this individual need to be determined by inferring the phase of the alleles. After phasing, the acetylator phenotype is assigned manually based on haplotypes (Supplementary Fig. 1). Phasing of alleles (i.e. haplotype determination) is laborious. Experimental methods exist, but are time-consuming and expensive. In most studies, computational statistical methods are used, such as the algorithm implemented in PHASE (Stephens *et al.*, 2001). However, methods for statistical determination of phase are computer intensive and require specific data formatting steps. The goal of the present work was to develop a web server that implements a supervised pattern recognition approach to infer NAT2 acetylator phenotype (slow, intermediate or rapid) directly from the observed combinations of *NAT2* SNPs, without taking the extra step of determining the haplotypes for each individual.

2 METHODS

The dataset used in this work was obtained from the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial of the National Cancer Institute (see Moslehi *et al.*, 2006 for details). Genotyping for six NAT2 SNPs (C282T, T341C, C481T, G590A, A803G and G857A) was performed using the TaqMan® (Applied Biosystems Inc., Carlsbad, CA, USA) kit. The acetylator phenotypes were assigned in our previous study based on the haplotypes determined from SNP genotyping data for each subject (Moslehi *et al.*, 2006). The dataset consists of 1377 subjects (see Supplementary Table 1 for details and ethnic makeup). Prediction of the acetylator phenotype from combinations of SNPs, as defined here, is a three-class classification problem that can be addressed using a supervised pattern recognition method. We used Support Vector Machine (SVM) as a method of choice (Vapnik, 1998). We constructed a three-class SVM predictor using the one-against-one approach which was shown to perform better than other approaches in multi-class SVMs (Hsu and Lin, 2002). We used SVM implemented in the LIBSVM package (Chang and Lin, 2003) with the linear kernel. Each *NAT2* SNP was encoded using a set of three mutually orthogonal binary vectors: homozygote for the most frequent allele (1,0,0), heterozygote for the most frequent allele (0,1,0) and homozygote for the least frequent allele (0,0,1). For a given subject corresponding vectors describing each of the six observed SNPs were concatenated together, resulting in a final binary feature vector of dimension 18. Thus, the SNP combination of each subject was described by 18 binary variables. We used a 7-fold cross-validation to test the SVM predictor of the acetylator phenotype. In this approach, the dataset is randomly partitioned into seven groups, each containing 1/7 of the dataset. At each cross-validation run, one group is removed and the predictor is trained on the remaining observations and tested on the removed group. The process is repeated seven times, so that each group is used for testing once. In order to assess different aspects of classification quality, we used the following performance measures: overall accuracy (ACC), sensitivity (SN)

*To whom correspondence should be addressed.

Table 1. The performance of NAT2PRED server

NAT2 phenotype	Sensitivity (SN)	Specificity (SP)
Rapid $n=84$	99.6% (93.4%)	100% (90.1%)
Intermediate $n=503$	100% (94.0%)	99.7% (95.4%)
Slow $n=790$	100% (92.5%)	100% (93.1%)

The total number of cases for a given phenotype is shown in a corresponding row name. The SVM penalty parameter C was set to 3 (an optimal value determined using a grid search). Numbers in parenthesis show the results of prediction based on non-synonymous SNPs.

for class i (SN_i) and specificity (SP) for class i (SP_i) (Baldi *et al.*, 2000):

$$\mathrm{ACC}=100\%\times\frac{\sum_i z[i,i]}{N} \tag{1}$$

$$\mathrm{SN}_i=100\%\times\frac{z[i,i]}{x[i]} \qquad \mathrm{SP}_i=100\%\times\frac{z[i,i]}{y[i]} \tag{2}$$

$$x[i]=\sum_j z[i,j], \qquad y[i]=\sum_j z[j,i] \tag{3}$$

where Z is a 3×3 confusion (contingency) matrix, in which an element z[i,j] represents the number of times objects from class i are predicted to be in class j; N is the total number of objects ($N=1377$ in this work).

3 RESULTS

The results of the cross-validation are shown in Table 1. If all six SNPs are used, the predictor of the NAT2 acetylator phenotype achieves a nearly perfect accuracy of 99.9% (Equation 1) and nearly perfect class-specific sensitivities and specificities (Equation 2) between 99.6 and 100%. Such a well-balanced performance is observed despite the highly unbalanced nature of the dataset, meaning that the number of subjects with the slow phenotype is almost an order of magnitude larger than that of subjects with the rapid phenotype. Importantly, individuals with the slow phenotype, who are at increased risk of developing tumors, are identified with 100% SN, meaning that no at-risk individuals are missed. If data on two synonymous SNPs (C282T and C481T) are removed and only non-synonymous SNPs are used, the accuracy of the prediction drops from 99.9 to 93.2%, with similar declines in SN and SP (Table 1). We therefore conclude that all six SNPs used in the present study are required to reliably assign the acetylator phenotype.

The web server implementation of the SVM predictor of the NAT2 acetylator phenotype was trained using the data on all 1377 subjects. It has a simple intuitive user interface .The user is asked to select a genotype for each of the six SNP loci using radio buttons (Supplementary Fig. 2). There are three possible genotypes for each SNP locus, which corresponds to three radio buttons per locus. After the genotype is selected, the user can click ‘Submit’ button and immediately obtain an inferred NAT2 acetylator phenotype. The output page displays the selected genotype and the probabilities of each of the three acetylator phenotypes (slow, intermediate and rapid) for these genotypes (Supplementary Fig. 3). The final prediction is the phenotype with the highest probability. There is also an option for a batch submission of genotypes for multiple individuals. Detailed instructions and information about the methodology and output format can be found by clicking the corresponding help hyperlink located on the input page. To the best of the authors’ knowledge, NAT2PRED is the only existing web server for inferring NAT2 acetylator phenotypes from genotyping data. NAT2PRED was developed on a dataset where majority of subjects are Caucasian (94%). However, the prediction model utilizes generally observed linkage disequilibrium between the six *NAT2* SNPs and can be applied to individuals from any ethnicity. The web server is publicly available at http://nat2pred.rit.albany.edu.

ACKNOWLEDGEMENTS

The authors thank Dr R. B. Hayes, Division of Cancer Epidemiology and Genetics and Drs C. Berg and P. Prorok, Division of Cancer Prevention, NCI, NIH, DHHS, the Screening Center investigators and staff of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial, Mr T. Riley and staff, Information Management Services, Inc., Ms B. O’Brien and staff, Westat, Inc. and Drs B. Kopp, W. Shao and staff, SAIC-Frederick. Most importantly, we acknowledge the PLCO study participants for their contributions to making this study possible.

Conflict of Interest: none declared.

REFERENCES

Baldi,P. *et al*. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Blum,M., *et al*. (1990) Human arylamine N-acetyltransferase genes: isolation, chromosomal localization, and functional expression. *DNA Cell Biol.*, **9**, 193–203.

Brockton,N. *et al*. (2000) N-acetyltransferase polymorphisms and colorectal cancer: a HuGE Review. *Am. J. Epidemiol.*, **151**, 846–861.

Chang,C.C. and Lin,C.J. (2003) LIBSVM: a library for support vector machines. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (last accessed date April 6, 2007).

Hein,D.W. *et al*. (2000) Molecular genetics and epidemiology of the NAT1 and NAT2 acetylation polymorphisms. *Cancer Epidemiol. Biomarkers Prev.*, **9**, 29–42.

Hein,D.W. (2006) N-acetyltransferase 2 genetic polymorphism: effects of carcinogen and haplotype on urinary bladder cancer risk. *Oncogene*, **25**, 1649–1658.

Hsu,C.W. and Lin,C.J. (2002) A comparison of methods for multi-class support vector machines. *IEEE Trans. Neural Netw.*, **13**, 415–425.

Moslehi,R. *et al*. (2006) Cigarette smoking, N-acetyltransferase genes and the risk of advanced colorectal adenoma. *Pharmacogenomics*, **7**, 819–829.

Potter,J. *et al*. (1999) Colorectal adenamatous and hyperplastic polyps: smoking and N-acetyltransferase 2 polymorphisms. *Cancer Epidemiol. Biomarkers Prev.*, **8**, 69–75.

Stephens,M. *et al*. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, **68**, 978–989.

Tiemersma,E. *et al*. (2004) Effect of SULT1A1 and NAT2 genetic polymorphism on the association between cigarette smoking and colorectal adenomas. *Int. J. Cancer*, **108**, 97–103.

Vapnik,V.N. (1998) *Statistical Learning Theory*. John Wiley & Sons, New York.