Learning classifiers from discretized expression quantitative trait loci

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

[1] CITIC, Universidad de Granada, Granada, Spain
[2] Instituto de Parasitología López Neyra, CSIC, Granada, Spain

Abstract. Expression quantitative trait loci are used as a tool to identify genetic causes of natural variation in gene expression. Only in a few cases the expression of a gene is controlled by a variant on a single marker. There is a plethora of different complexity levels of interaction effects within markers, within genes and between marker and genes. This complexity challenges biostatisticians and bioinformatitians every day and makes findings difficult to appear. As a way to simplify analysis and better control confounders, we tried a new approach for association analysis between genotypes and expression data. We pursued to understand whether discretization of expression data can be useful in genome-transcriptome association analyses. By discretizing the dependent variable, algorithms for learning classifiers from data as well as performing block selection were used to help understanding the relationship between the expression of a gene and genetic markers. We present the results of a first set of studies in which we used this approach to detect new possible causes of expression variation of DRB5, a gene playing an important role within the immune system. A supplementary website including a link to the software with the method implemented can be found at http://bios.ugr.es/classDRB5.

Keywords: SNP, eQTL, HapMap, gene expression microarray

1 Introduction

Association between genotypes and mRNA transcript levels may help elucidating genetic basis of complex diseases by analyzing whenever genetic variants affect gene expression. A variant affecting a disease may be found also in association with a gene expression, i.e., it may be an expression quantitative trait loci (eQTL). However, it is not straightforward to understand whether it may truly alter gene transcription or splicing or just being in linkage disequilibrium (LD) with the real cause [23]. Moreover, because of small sample sizes and limited computational resources, regression models using a few input variables have given results hardly reproducible and most succesful association analyses only have tested single polymorphic loci (SNP) against the expression of a gene instead of considering more than one SNP at a time [7, 23, 21]. We have used a different approach to measure SNP-expression data association which relies on a pre-discretization of expression data as a way to simplify input data and improve performance compared with standard regression models. Discretization of gene expression data is commonly performed when they are used as the input variables to predict different phenotypes such as cellular classification in cancer [15, 18, 24]. With this simplification, we are able to use data to learn a classifier, i.e. a model that relates how input variables, the SNPs, and their interaction, affect a discrete output variable, high or low gene expression in our assumption of only two bins. If this was the case, other more complex analyses, such as considering multiple SNPs at a time or haplotyping analysis [1] could reduce computational and statistical complexity becoming both more affordable and powerful. Therefore, we may use different approaches to learn classifiers and, as they make different assumptions, shed light about the main features of the data analyzed and thus about different interaction patterns between genes and regulatory proteins affecting their expression and about their association with SNPs within a block of high LD.

We focused on a gene HLA-DRB5 (DRB5). DRB5 is one of the genes that encode β chains for the DR HLA class II receptor. The HLA genes are located on the short arm of chromosome 6 and are organized in three regions: MHC class I, MHC class II and MHC class III. HLA class II genes encode glycoproteins expressed primarily on antigen-presenting cells where they present processed antigenic peptides to CD4+ T cells. The DR β chain is encoded by 4 gens DRB1, 3, 4, and 5. There are also other pseudogenes that do not produce a protein: DRB2, 7, 8, and 9. Not everybody has a copy of each gene or pseudogene. There are 5 different haplotypes with different combinations of genes. DRB5 is only present in DR51 haplotype. This haplotype has been associated with immune related diseases susceptibility. In particular, the DRB5*0101- DRB1*1501-DQA1*0102-DQB1*0602 haplotype has been associated with MS in the North European population [5]. These alleles are almost always present together in this population, making it impossible to distinguish the primary association. DRB alleles have different structural properties for antigen presentation according to their amino acid sequence. This points out to a specific antigen presentation as the pathogenic mechanism through this does not fully explain the disease association. The description of polymorphisms that alter HLA gene expression [22] and the identification of several cis-acting genetic variants on expression of HLA class II genes [19, 4], makes it possible that association of HLA class II polymorphisms with MS may be related to the levels of gene expression to the same or a greater extent than restriction of antigen response. In fact the ability to induce active experimental autoimmune encephalomyelitis (EAE), an animal model for MS disease, was increased in animals expressing higher levels of DRB5*01:01, pointing to a role of the levels of expression of this gene in susceptibility [13].

We chose this gene because the expression pattern in the two data sets analyzed showed two statistical modes and it was easily translated to a binarized pattern. In Section 2 we describe the methods we used. Data sets and their attributes are described in Section 3. Results appear in Section 4 and conclusions and future work can be read in Section 5.

2 Method

The starting point of our method is a binarized variable which represent the level of expression of a gene in an individual. We do not focus in the discretization problem in this work. For the gene we used, DRB5, discretization was a straightforward task, as we could assume a density function with two non-overlapping normal distributions. Figure S1 (a) and (b) show histograms corresponding to DRB5 expression levels in the two data sets (CEU and YRI, respectively) used.

This study focused on only one gene with a well-known expression behavior as our purpose was to understand whether discretization may be a valid method to increase power. Therefore we were able to explore the genotypes at the whole chromosome so that both cis and trans within a chromosome eQTLs could be detected. As a way to reduce computational time when using this method with many genes only cis eQTLs could be first analysed.

2.1 Classification algorithms

A classification learning algorithm starts from a previously given set of training data sets D and try to learn a function $f : X \rightarrow Y$, where X is the set of input variables (i.e. a subset of SNPs) and Y is the output variable (i.e. the expression of the gene: low or high).

In order to deeply understand the population bias [26] and shed light about genetic causes of differential expression, we chose three state-of-the-art and computationally affordable algorithms to learn classification models for predicting the expression level of the DRB5 gene. The first approach was the Naive Bayes (NB) classifier, which is a simple probabilistic classifier which works under the assumption that all input variables are conditionally independent given the output variable. The second approach was classification trees, in which the data set is divided in structured hypercubes of those individuals sharing values. We chose its most competitive algorithm, C4.5 [17], called J48 in an open-source version implemented in Weka [8]. For the last approach, we chose support vector machines (SVM), in which the input variables are transformed in a higher-dimension space so that a classifier is learned from the set of transformed variables by using a kernel function. We chose the default implementation of support vector machines in Weka, which uses the Sequential Minimal Optimization [16] as the learning algorithm. One advantage of the first two approaches is that they build white-box models, i.e., models are directly readable and interpretable by human experts.

2.2 Block processing

As a natural way to select input attributes in genome-wide data, we may group them in blocks of low recombination, i.e., SNPs with high LD among them. Thus, we grouped SNPs by using a common approach based on pairwise computations of confident intervals of LD [6]. As LD is known to be higher in individuals with European ancestries than from Africa [11], we made blocks by using CEU, the data set of individuals with European ancestries (see Section 3) and used those blocks to group SNPs in YRI, the data set of individuals with African ancestries. Figure S2 shows the average number of SNPs by block in the data set used. Given a block, a classifier with only those SNPs in that block as input variables was learned.

Figure 1 shows a flowchart with all the steps followed to conduct this study.

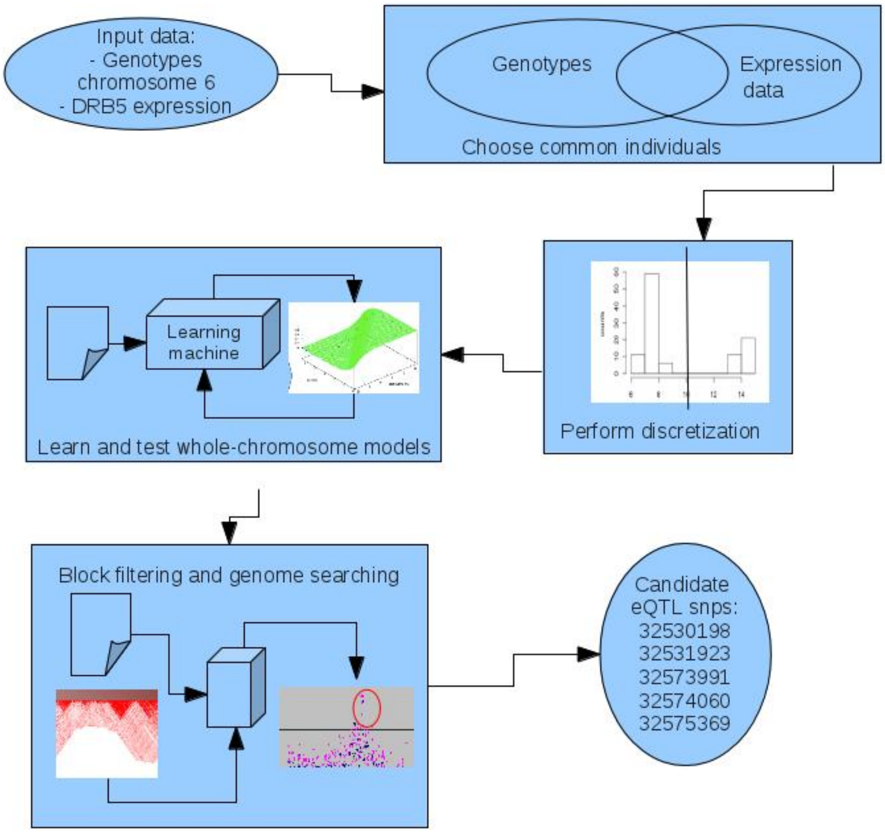

Fig. 1. Flowchart showing the steps followed in this study in order to obtain a minimal set of candidate eQTLs for expression of gene DRB5.

3 Data sets used

As data sets we used all (113) the unrelated individuals (parents) from CEPH (CEU), an Utah (USA) data set of individuals with Northern and Western European and all (113) the unrelated individuals (parents) from a data set (YRI) of the Yoruba (Nigeria) population. Genotypes in chromosome 6 (where gene DRB5 belongs to) of these individuals passing quality control (6593) [10] were chosen from the third phase of HapMap project [9] to avoid large amounts of missing data, as in the other phases not all these individuals were genotyped. Haplotypes and missing information were inferred by using familial information and the IMPUTE2 algorithm [12] and these data were downloaded from the HapMap project website (http://www.hapmap.org).

Gene expression data was obtained for only 108 individuals from CEU and 107 from Yoruba, so that we only used genotypes of these individuals as well. Expression data came from the RNA of lymphoblastoid cell lines. Details about the procedure followed to obtain expression data, including raw expression data normalization, population stratification correction and correction for known and unknown factors are described by Stranger et al. [21]. Gene DRB5 was found lowly expressed in 63.5% of the CEU data set and 70% of the YRI data set.

SNPs in chromosome 6 were grouped in 345 non-overlapping blocks of low recombination [6] which were learned from the CEU data set.

DRB5 is a gene coded between physical positions 32593098 and 32606042 in assembly NCBI36/hg18 or between 32485120 and 32498064 in assembly GRCh37/hg19. 6 SNPs has been genotyped in HapMap 3 within the gene DNA region. These SNPs belong to block 223. Table S1 at the supplementary material shows the SNPs within the block. Those in bold correspond to SNPs within the gene.

4 Results

In this section we describe the results we obtained when using all the SNPs to learn classifiers (Subsection 4.1) and when using only those SNPs within a block and models with single SNPs (Subsection 4.2). Because LD is lower in African populations than in European populations, we only show results from YRI population as the number of candidate eQTL SNPs among all found in association with the DRB5 expression is lower than if we used CEU data set. However, very similar results in terms of predictive capacity of the different approaches used were found when using the CEU data set (data not shown).

4.1 Whole models

In a first step using all SNPs at a time, we studied the extent to which the expression level of DRB5 was fully controlled by the SNPs chosen and whether redundant or noisy variables could affect the overall performance. Thus, we learned classifiers using all SNPs in the data sets and the three different learning machines: NB, C4.5 and SVM. To measure generalization capacity of the learned models, we used 10-fold cross-validation (10-cv) [14], so that results are the mean between 10 disjoint test data subset of 1/10 of the original size when the model was learned with the remaining individuals. The area under the ROC curve (AUC) was very high when using NB and reached its maximum (perfect classification) for C4.5. See Figures S3 (a), (b) and (c) for the ROC curves obtained when using NB, C4.5 and SVM respectively. See Table 1 for classification accuracy in percentage (column 2), AUC (column 3) ans relative absolute error (column 5). For comparative purposes, we also showed results when using original continuous expression data and two state-of-the-art algorithms to learn regression models: SVM-reg [20], which are based on *support vector models* and *kernel methods*, as its supervised classification counterpart; and Gaussian processes [25], which is a Bayesian approach that employs Gaussian process priors over regression functions to improve their generalization capacity. The relative absolute error (column 5) is defined for both classification and regression models and it shows how all the classification models used outperform all the regression models.

Table 1. Generalization capacity of different learning machines using all SNPs in the data sets as input variables.

Predictive Models	Accuracy	Area under Roc	Root Mean Square Error	Relat. Abs. Error
NB	89.7%	0.89	–	22.3%
C4.5	99.0%	1.00	–	1.6%
SVM	79.5%	0.66	–	48.8%
SVM-Reg.	–	–	3.22	72.3%
Gaussian Proc.	–	–	2.1	55.1%

An outstanding result is given by $C4.5$ and YRI data set, with 99% accuracy and 1.0 AUC. The results become even more interesting due to the white-box nature of $C4.5$ and its pruning feature so that the subset of variables not pruned in the tree-like model can help biomedical researchers to understand SNPs regulation the expression of DRB5. On the contrary, classifiers based on SVM and regression models had a low predictive capacity even when is well-known that expression regulation of DRB5 is controlled by genetic variants in chromosome 6 tagged by some of the SNPs from the genotype array used.

4.2 Block-based approach

In a second step using blocks, we tried to understand the consequences on classification performance of reducing the number of input variables by using block selection, a biologically-inspired feature selection method. Would accuracy keep as higher as when all SNPs were used?

For each data set we used NB, SVM and C4.5 algorithms to compute AUC and accuracy in 10-cv using as input variables only those SNPs within a block.

We repeated this process for each block out of the 345 blocks which chromosome 6 was divided by.

Figure S4 shows that AUCs, which reached its maximum using C4.5 and all SNPs, is also maximum by using only single blocks with SVM and NB for several blocks. Moreover, values seem to follow an ascending function from the chromosome extremes to blocks 223 to 224 in which AUC=1 for the three learning machines used (results for C4.5 are not shown). A simple explanation is that DRB5 is coded in block 223 and SNPs in this block capture allelic mutations in DRB5 that avoid its transcription or may be cis-acting regulatory factors. In order to understand how many SNPs within block 223 or others with high performance are tag SNPs and how they interact, we first have computed AUC and accuracy by using single SNPs as input variables. Figure 2 shows AUC for NB using blocks (pink) and the SNP within a block with maximum AUC (blue). Similar results are obtained with the other classifiers. Table 2 (a) shows those single SNPs and the blocks they belong to when AUC reached its maximum for all the algorithms used. Figure S5 (a) shows how perfect classification using single SNPs means perfect association between expression level (red is high expression and blue means low expression) and genotypes (from left to right: homozygotic wild type, heterozygotic and homozygotic mutant allele) using any of these SNPs.

High performing in adjacent blocks may be due to high LD, something that is supported by the descending-with-distance pattern showed in Figure 2.

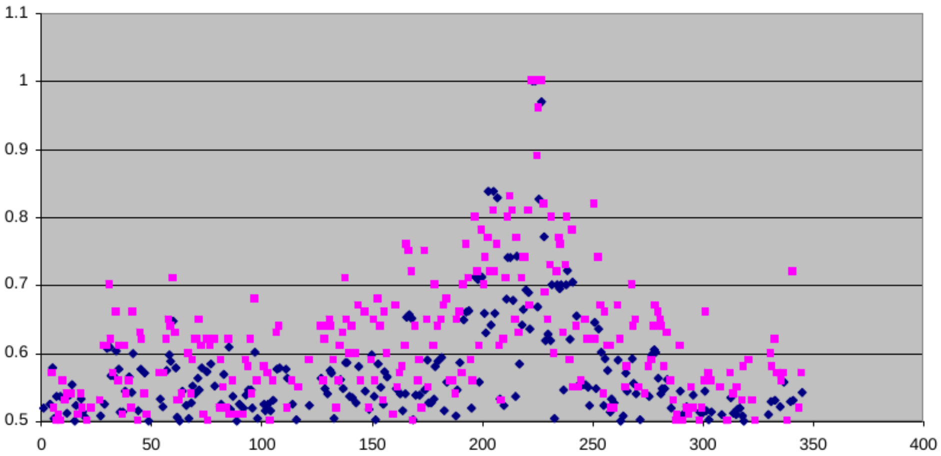

Fig. 2. YRI data set. AUC (y-axis) from NB using single blocks (pink) and single SNPs (blue) with maximum classification performance (blue) with 10-cv. First SNP for the block analyses or the SNP for analyses using single SNPs are shown on x-axis.

Table 2. List of SNPs with highest performance (AUC=1) and their respective blocks.

SNP (Physical position assembly GRCh37/hg19)	Block number
32530198	223
32531923	223
32573991	224
32574060	224
32575369	224

5 Conclusions

In this work we have shown how by discretizing gene expression, classification learning machines can be used to learn and test complex models made up of hundred or thousand input variables which are robust to redundancy and noisy variables. Results obtained for gene DRB5 have extensively been confirmed by the more standard regression models, as they only require mono-variate models to explain DRB5 expression. The levels of DR gene expression could condition the type of immune response. The high expression of DRB5 gene could affect directly the concentration of peptide-MHC complex on the antigen presenting cell (APC) and in turn, to affect the duration and specificity of the T cell-TCR with APC-HLA molecules interaction. It has been shown that the immunological synapse strength in the interaction between the antigenpresenting cells (APC) and the T cell determines the fate of T cells into Th1 or Th2 types [3]. The stronger TCR signal favors Th1 differentiation and this is dependent on the potency of TCR/peptide-MHC interactions, density of peptide-MHC complexes including co-stimulatory molecules and the duration of T cell-APC contacts. Therefore, the higher or lower expression of the different HLA molecules, opens a spectrum of possible combinations with specific structure receptors and levels of co-stimulatory molecules that would determine the magnitude and quality of the T cell response and the type of fate decision made by peripheral T cells [2].

We have shown how several classification algorithms are robust to redundancy and multi-variate classifiers are still able to keep predictive accuracy. Some of the classification approaches have revealed to be very helpful for biomedical researchers, as they have learned white-box models easily interpretable by human experts. Feature selection of SNPs based on LD criterion has helped to identify SNPs that may be candidate eQTLs in the predictive models. Because of their low computational complexity to the number of input variables, we have been able to use very robust classification algorithms under different approaches with all the SNPs within chromosome 6. These conclusions are more difficult to obtain when using regression models, as the larger complexity of a regression model compared with a classifier translates into a reduction in robustness to redundancy and therefore in generalization capacity and interpretability.

Acknowledgment

The authors were supported by the Spanish Research Program under project TIN2010-20900-C04-1, the Andalusian Research Program under project P08-TIC-03717 and the European Regional Development Fund (ERDF).

References

1. Abad-Grau, M., Medina-Medina, N., Montes-Soldado, R., Matesanz, F., Bafna, V.: Sample reproducibility of genetic association using different multimarker tdts in genome-wide association studies: Characterization and a new approach. PLoS ONE 7(2), 29613 (2012)
2. Alcina, A., Abad-Grau, M., Fedetz, M., Izquierdo, G., Lucas, M., Fernndez, O., Ndagire, D., Catal-Rabasa, A., Ruiz, A., Gayn, J., Delgado, C., Arnal, C., Matesanz, F.: Multiple sclerosis risk variant hla-drb1*1501 associates with high expression of drb1 gene in different human populations. PLoS One 7(1), e29819 (2012)
3. Corse, E., RA, R.A.G., Allison, J.P.: Strength of tcr-peptide/mhc interactions and in vivo t cell responses. Journal of Immunology 186, 5039–5045 (2011)
4. Dixon, A., Liang, L., Moffatt, M.F., Chen, W., et al., S.H.: A genome-wide association study of global gene expression. Nature Genetics 39, 1202–1207 (2007)
5. Fogdell, A., Hillert, J., Sachs, C., Olerup, O.: The multiple sclerosis- and narcolepsy-associated hla class ii haplotype includes the drb5*0101 allele. Tissue Antigens 46, 333–336 (1995)
6. Gabriel, S., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E., Daly, M., Altshuler, D.: The structure of haplotype blocks in the human genome. Science 296, 2225–9 (2002)
7. Gat-Viks, I., Meller, R., Kupiec, M., Shamir, R.: Understanding gene sequence variation in the context of transcription regulation in yeast. PLoS Genetics 6(1), e1000800 (2010)
8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. In: SIGKDD Explorations. vol. 11, p. 1 (2009)
9. HapMap-Consortium, T.I.: The international hapmap project. Nature 426, 789–796 (2003)
10. HapMap-Consortium, T.I.: Integrating common and rare genetic variation in diverse human populations. Nature 467(7311), 52–58 (Sep 2010), http://dx.doi.org/10.1038/nature09298
11. Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A., , Cox, D.R.: Whole-genome patterns of common dna variation in three human populations. Science 18, 1072–79 (2005)
12. Howie, B.N., Donnelly, P., Marchini, J.: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics 5(6), e1000529 (2009)
13. JA, J.Q., Huh, J., M, M.B., Yao, K., N, N.I., Bryant, M., Kawamura, K., Pinilla, C., McFarland, H., Martin, R., Ito, K.: Myelin basic protein-specific tcr/hla-drb5*01:01 transgenic mice support the etiologic role of drb5*01:01 in multiple sclerosis. Journal of Immunology 189(6), 2897–908 (2012)

14. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 15th International Joint Conference on Artificial Intelligence. pp. 114–119 (1995)
15. Liu, H., Li, J., Wong, L.: A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome Informatics 13, 51–60 (2002)
16. Platt, J.C.: Sequential minimal optimization: A fast algorithm for training support vector machines. Tech. rep., Advances in kernel methods - Support Vector Learning (1998)
17. Quinlan, R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc. (1993)
18. Ross, M.E., Zhou, X., Song, G., Shurtleff, S.A., Girtman, K., Williams, W.K., Liu, H.C., Mahfouz, R., Raimondi, S.C., Lenny, N., Patel, A., Downing, J.R.: Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. Blood 102(8), 2951–2959 (2003)
19. Schadt, E., Molony, C., Chudin, E., Hao, K., et al., X.Y.: Mapping the genetic architecture of gene expression in human liver. PLoS Biology 2008, 6:e107 (2008)
20. Shevade, S.K., Keerthi, S.S., Bhattacharyya, C., Murthy, K.R.K.: Improvements to the smo algorithm for svm regression. IEEE Trans. Neural Netw. Learning Syst. 11(5), 1188–1193 (2000)
21. Stranger, B.E., Montgomery, S.B., Dimas, A.S., Parts, L., Stegle, O., Ingle, C.E., Sekowska, M., Smith, G.D., Evans, D., Gutierrez-Arcelus, M., Price, A., Raj, T., Nisbett, J., Nica, A.C., Beazley, C., Durbin, R., Deloukas, P., Dermitzakis, E.T.: Patterns of cis regulatory variation in diverse human populations. PLoS Genetics 8(4), e1002639 (2012)
22. Vincent, R., P, P.L., Gongora, C., Papa, I., et al., J.C.J.: Quantitative analysis of the expression of the hla-drb genes at the transcriptional level by competitive polymerase chain reaction. Journal of Immunology 156, 603–610 (1996)
23. Wallace, C., Rotival, M., Cooper, J.D., Rice, C.M., Yang, J.H., McNeill, M., Smyth, D.J., Niblett, D., Cambien, F., Consortium, C., Tiret, L., Todd, J.A., DG, D.G.C., Blankenberg, S.: Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. Human Molecular Genetics 21(12), 2815–24 (2012)
24. Wanga, Y., Tetko, I.V., Hall, M.A., Frank, E., Facius, A., Mayer, K.F., Mewes, H.W.: Gene selection from microarray data for cancer classificationa machine learning approach. Computational Biology and Chemistry 29(1), 37–46 (2004)
25. Williams, C., Rasmussen, C.: Gaussian Processes for Regression. In: Advances in Neural Information Processing Systems 8. vol. 8, pp. 514–520 (1996)
26. Wilson, D.R., Martínez, T.R.: Bias and the probability of generalization. In: Proceedings of the International Conference on Intelligent Information Systems (IIS). pp. 108–114 (1997)