Application of Classification Algorithms on IDDM Rat Data

Rainer Schmidt[1], Heike Weiss[2], and Georg Fuellen[1]

[1] Institute for Biostatistics and Informatics in Medicine and Aging Research
[2] Institute for Medical Biochemistry and Molecular Biology,
University of Rostock, Germany
{rainer.schmidt,heike.weiss,georg.fuellen}@uni-rostock.de

Abstract. In our study, we intend to investigate the mechanism of tolerance induction by the modulatory anti CD4 monoclonal antibody RIB 5/2 in insulin dependent diabetes mellitus rats. The aim of this investigation is to identify the key mechanisms of immune tolerance on the level of T cell, cytokine, and chemokine biomarkers in the blood, lymphatic organs, and pancreas. Additionally, it should be possible to define good biomarkers of autoimmunity and tolerance for prediction of diabetes onset. We mainly applied decision trees and later on some other classification algorithms on a rather small data set. Unfortunately, the results are not significant but are good enough to satisfy our biological partners.

Keywords: Insulin dependent diabetes mellitus, Bioinformatics, Machine Learning.

1 Introduction

Type 1 diabetes is an autoimmune disease in which beta cells are exclusively destroyed by the interaction of antigen presenting cells, T cells, and environmental triggers such as nutrients and viral infection [1, 2].

There are two major challenges for prediction and diagnosis of this disease. First, though the analysis of various beta cell autoantibodies and beta cell specific T cells allows a good risk assessment for the progression of autoimmunity, biomarkers related to mechanisms of T cell mediated beta cell destruction and induction of self-tolerance are missing. Second, intervention strategies to block beta cell autoimmunity are not fully understood.

The IDDM (insulin dependent diabetes mellitus) rat is an animal model of spontaneous autoimmune diabetes which is characterized by a fulminant T cell mediated beta cell destruction leading to a full diabetic syndrome in 60 % of the animals around day 60. The narrow time range of islet infiltration between day 40 and day 50 makes this model a valuable tool to study strategies and mechanisms for induction of immune tolerance. Induction of immune tolerance is a promising approach to halt autoimmunity in type 1 diabetes. Anti CD3 antibodies and vaccination with modified

P. Perner (Ed.): ICDM 2012, LNAI 7377, pp. 1–8, 2012.

beta cell antigens such as insulin, GAD65, and hsp60 could block autoimmunity and induce self-tolerance in animal models of autoimmune diabetes [3].

These strategies, however, still show limitations that hamper translation into routine clinical use. First, the mechanisms of T cell modulation are still unclear in particular for transition from temporary immune suppression to induction of permanent self-tolerance. Second, despite development of humanized and aglykosylated anti CD3 antibodies the side effects remain severe and raise ethical concerns for treatment of young type 1 diabetes patients.

The intention of our project is

- To elucidate the mechanisms of the modulating anti CD4 antibody RIB5/2 on prevention of autoimmune destruction of beta cells in the IDDM rat model.
- To analyse immune cell (bio-) markers in peripheral blood during progression of autoimmunity and/or induction of self-tolerance.

2 Background and Research Status

Beta cell destruction in type 1 diabetes is a complex process comprising a network between beta cells, antigen presenting cells, autoagressive T cells, and environmental triggers. Beta cells that are under assault are not passive bystanders, but actively participate in their own destruction process [4, 5]. Overall, many of the cytokine- and virus-induced effects involved in inhibition of beta cell function and survival are regulated at the transcriptional and posttranscriptional/translational level [6]. T-cells modulate the autoimmune process and autoreactive T-cells can transfer diseases [7]. Thus, immune intervention during the prodromal phase or at the onset of overt diabetes will affect the balance between autoreactive and regulatory T cells. Currently it is possible to identify ß-cell-specific autoreactive T-cells using standard in vitro proliferation and tetramer assays, but these cell types could also be detected in healthy individuals [8]. Although the analysis of autoantibodies allows an assessment of risk for type 1 diabetes, it is still impossible to draw conclusions about T cell function in the local lymphatic compartment of the pancreas. Notably, there is an extensive knowledge upon activation of T cells and upon induction of self-tolerance on the molecular level of gene expression biomarkers. We hypothesize that biomarkers must be analyzed in a dynamic manner because they shall have specific predictive values for development of autoimmunity at different stages of autoimmunity.

The analysis of gene expression patterns might help to distinguish between T1DM affected subjects and healthy animals at an early stage. In a first experiment, we could demonstrate that analysis of selected genes of T cell differentiation, T cell function, and cytokine expression in whole blood cells at an early prediabetic stage (after 45 days of live), the RT6 T cell proliferation gene was most decisive for diabetes onset in the IDDM rat followed by selectin and neuropilin at the stage of islet infiltration (after 50 days), and IL-4 during progression of beta cell destruction (after 55 days).

3 Data

Several experiments were performed and statistically evaluated. In one of them, for example, it could be shown that the treatment of prediabetic IDDM rats with antibody RIB 5/2 significantly reduces diabetes incidence (from 60% to 11%, see figure 1).

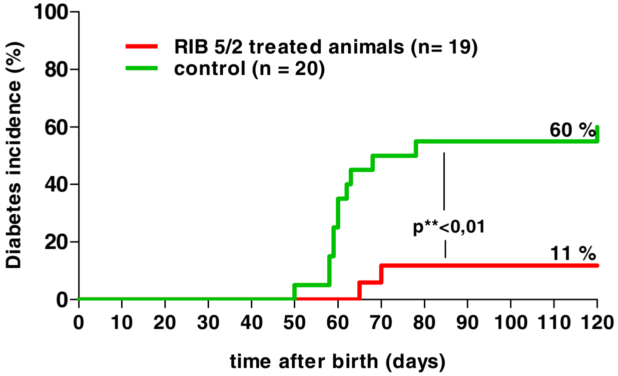

Fig. 1. Protective effect of RIB/2 CD4 modulation on diabetes incidence and age of manifestation (p<0.01, chi-test)

For recent experiments,data from twelve rats were available. They were monitored for gene expression data in blood immune cells for functional gene clusters on the days 30, 35, 40, 45, 50, 55, 60, 65, 70, 80, and 90 of their life. However, just the days between 45 and 60 are assumed to be important of the prediction whether a rat will develop diabetes or not. Six of the twelve rats developed diabetes, three did not, and another three rats (background strain) were diabetes resistant because of the way they had been bred. Unfortunately, due to problems of the measurement facilities the data quality is rather poor. Many data are missing and some are obviously incorrect, especially for the early and the late measurement time points. However, as mentioned above, the most important measurement time points are in the middle. So, for some measurement time points, data from just eleven of the twelve rats were used.

4 Experimental Results

In the experiments data of the following measurement time points were used: 45, 50, 55, and 60 days of life. The attributes are eighteen preselected genes and biomarkers. The class labels are “diabetes”, “no diabetes”, and “background strain”.

Since we wanted to get attributes that are most decisive for the classification, we applied decision trees, which do not just provide the most decisive attributes but also their decisive values. The C4.5 decision tree algorithm, which was originally developed by Ross Quinlan [9], was applied in form of its J48 implementation in the

WEKA environment [10]. Later on, we also applied other classification algorithms that are provided in WEKA, like “random forest”, for example.

The tree for day 50 is depicted in figure 2 and states the following. If the gene expression value of selectin is bigger than 2.14 a rat probably belongs to the background strain, otherwise if the gene expression value of neuropilin is bigger than 0.63 a rat probably does not develop diabetes, otherwise it probably develops diabetes.

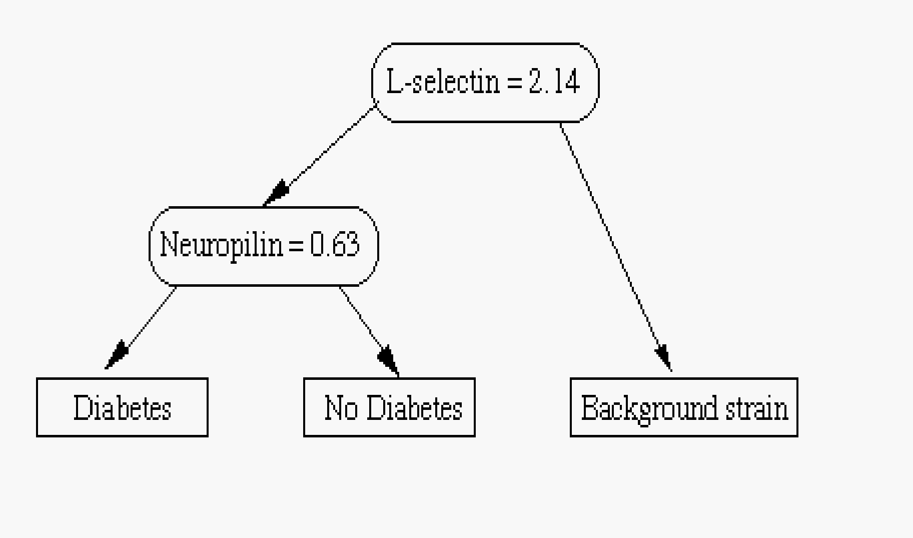

Fig. 2. Decision tree for day 50

The results for the days 45, 50, and 55 are depicted in figure 3. There are three trees depicted, the left one is for day 45, the right one is for day 55. The tree for day 50 (the same as in figure 2) is depicted in the middle.

At the beginning of infiltration (day 45) the RT6 gene expression, responsible for the correct thymic development of T-cells, may decide whether autoimmunity could develop. At a stage of of islet infiltration (day 50) selectin and neuropilin gene expression decides whether primed T-cells will infiltrate the endocrine pancreas for beta cell destruction. During progression of beta cell destruction (day 55) IL-4 as a T cell stimulating cytokine is crucial for the progression of beta cell infiltration.

5 Validation

For the IDDM rat model we have started to calculate relative risk coefficients for development for diabetes. Though the data set is very small, the biologists that are involved in the project are very happy with the results and can explain them (caption of figure 1). However, because of the extremely small data set (twelve rats), the set was not split into a learning and a test set. The trees are computed on the training set.

So, next Information Gain [10] was considered, on which decision trees are based. WEKA provides them as “attribute selection”. Usually, the values are between 0 and 1. In three of four trees the decision was obvious. For day 45, for example, the value of rt6a is 0.811, whereas the values of all other attributes are 0. Just, for day 50 the decison is obvious but the whole situation is not completely clear, because the Information Gain values are 0.959 for l-selection and 0.593 for il-4 and for neuropilin. Furthermore, in the tree neuropilin is used to separate between “diabetis” and “no diabetis”. So, the background strain was excluded and Information Gain was used just to classify “diabetis” and “no diabetis”, with the result that neuropilin was the first choice.

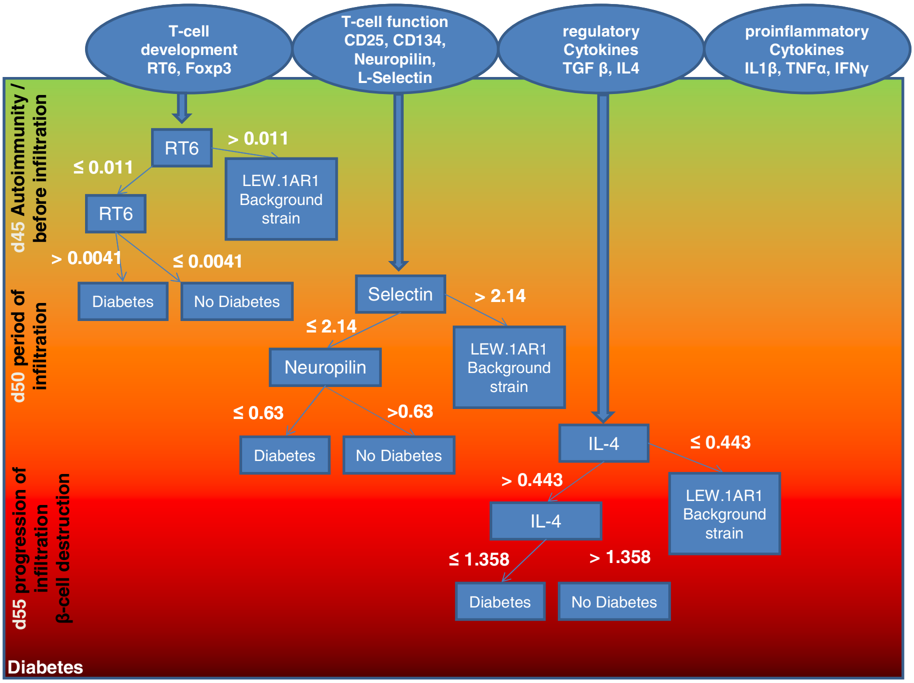
T-cell development RT6, Foxp3
T-cell function CD25, CD134, Neuropilin, L-Selectin
regulatory Cytokines TGF β, IL4
proinflammatory Cytokines IL1β, TNFα, IFNγ
d45 Autoimmunity / before infiltration
d50 period of infiltration
d55 progression of infiltration β-cell destruction
Diabetes
RT6
> 0.011
≤ 0.011
LEW.1AR1 Background strain
RT6
> 0.0041
≤ 0.0041
Diabetes
No Diabetes
Selectin
> 2.14
≤ 2.14
Neuropilin
LEW.1AR1 Background strain
≤ 0.63
>0.63
Diabetes
No Diabetes
IL-4
≤ 0.443
> 0.443
IL-4
LEW.1AR1 Background strain
≤ 1.358
> 1.358
Diabetes
No Diabetes

Fig. 3. Relative gene expression levels analysed with the C4.5 algorithm. The numbers heading the arrows indicate the threshold values of gene expression normalized to GAPDH quantified by real-time RT-PCR analysis.

Afterwards some standard classification methods provided by WEKA were applied (naïve bayes [11], nearest neighbor [12], random forest [13], J48, and support vector machines [14]). Except for the decision tree algorithm J48 these methods show just the classification results but they do not show which attributes have been used for the classification. In table 1 results are shown just for day 50 as an example. An inner cross validation is provided by WEKA. Because of the small size of the data set 3-fold cross validation was applied instead of the usual 10-fold cross validation. First, the classification algorithms were applied on the whole data sets (table 1) and secondly on the data sets without background strain (table 2). However, the differences are very small.

Table 1. Accuracy and Area Under the Curve for day 50 for the complete data set

Method	Accuracy (%)	AUC
Naïve Bayes	58.3	0.52
Nearest Neighbor	75	0.75
Random Forest	66.7	0.80
J48	66.7	0.76
SVM	58.3	0.65

Table 2. Accuracy and Area Under the Curve for day 50 for the data set without background strain

Method	Accuracy (%)	AUC
Naïve Bayes	55.6	0.42
Nearest Neighbor	77.8	0.67
Random Forest	66.7	0.53
J48	66.7	0.61
SVM	55.6	0.50

6 Discussion

The application of Machine Learning (classification) methods has become populare in bioinformatics. This is already reflected at the ICDM conferencerences (e.g. [15,16]). In our application, the analysis of gene expression patterns might help to distinguish between T1DM affected subjects and healthy animals at an early stage. In a first experiment, we could demonstrate that analysis of selected genes of T cell differentiation, T cell function, and cytokine expression in whole blood cells at an early prediabetic stage (after 45 days of live), the RT6 T cell proliferation gene was most decisive for diabetes onset in the IDDM rat followed by selectin and neuropilin at the stage of islet infiltration (after 50 days), and IL-4 during progression of beta cell destruction (after 55 days).

However, so far the data set is very small and, probably because of poor data quality, the cross-validated classification results are not significant (see tables 1 and 2). Nevertheless, the generated decision trees perform well, certainly just on the training set, but nearly all of them can be very well explained by the biochemical experts.

So, because of the small size of the data set, we tried to breed some more specific rats. Unfortunately, because of a virus in the rat laboratory this was just partly successful. Furthermore, since the data quality was poor, we applied another measurment facility. We got a better data quality for just eight new rats. However, both data sets should not be joined together. Because of different measurement facilities they are not compatible with each other.

The eight new cases belong to just two classes. Six rats developed diabetes, the other two ones did not. With such a data set the application of decision trees does not seem to be reasonable, because there is a big chance that one attribute can be found that might be sufficient to split away the two rats from the remaining six diabetes rats.

Actually, the situation is even worse. Most attributes can be used to distinguish between the two classes. Since all attributes we found in our earlier experiments (see figure 3) are among them, this new data set supports our findings. However, this support is rather weak, because for these new data many alternative attributes can also be used to separate between the two classes.

Unfortunately, the breeding and the data collection of these specific rats is expensive and time consuming. Furthermore, sometimes the breeding may even fail (see above).

However, our results (especially the decision trees) are not just good enough to satisfy our biological partners but also to get the funding, so that we are going to start another breeding attempt.

References

1. Akerblom, H.K., Vaarala, O., Hyoty, H., Ilonen, J., Knip, M.: Environmental factors in the etiology of type 1 diabetes. Am. J. Med. Genet. 115, 18–29 (2002)
2. Jun, H.S., Yoon, J.W.: A new look at viruses in type 1 diabetes. Diabetes Metab. Res. Rev. 19, 8–31 (2003)
3. Ludvigsson, J., Faresjo, M., Hjorth, M., et al.: GAD treatment and insulin secretion in recent-onset type 1 diabetes. N. Engl. J. Med. 359, 1909–1920 (2008)
4. D'Hertog, W., Overbergh, L., Lage, K., et al.: Proteomics analysis of cytokine-induced dysfunction and death in insulin-producing INS-1E cells: new insights into the pathways involved. Mol. Cell Proteomics 6(21), 80–99 (2007)
5. Rasschaert, J., Liu, D., Kutlu, B., Cardozo, A.K., Kruhoffer, M., ØRntoft, T.F., Eizirik, D.L.: Global profiling of double stranded RNA- and IFN-gamma-induced genes in rat pancreatic beta cells. Diabetologia 46, 1641–1657 (2003)
6. Gysemans, C., Callewaert, H., Overbergh, L., Mathieu, C.: Cytokine signalling in the beta-cell: a dual role for IFNgamma. Biochem. Soc. Trans. 36, 328–333 (2008)
7. Lampeter, E.F., McCann, S.R., Kolb, H.: Transfer of diabetes type 1 by bone-marrow transplantation. Lancet 351, 568–569 (1998)
8. Schloot, N.C., Roep, B.O., Wegmann, D.R., Yu, L., Wang, T.B., Eisenbarth, G.S.: T-cell reactivity to GAD65 peptide sequences shared with coxsackie virus protein in recent-onset IDDM, post-onset IDDM patients and control subjects. Diabetologia 40, 332–338 (1997)
9. Quinlan, J.R.: C4.5 Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993)
10. Hall, M., et al.: The WEKA data mining software: An update. SIGKDD Explorations 11(1), 10–18 (2009)
11. Gan, Z., Chow, T.W., Huang, D.: Effective Gene Selection Method Using Bayesian Discriminant Based Criterion and Genetic Algorithms. Journal of Signal Processing Systems 50, 293–304 (2008)

12. Cost, S., Salzberg, S.: A weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. Machine Learning 10(1), 57–78 (1993)
13. Breiman, L.: Random Forest. Machine Learning 45(1), 5–32 (2001)
14. Platt, J.: Avances in Large Margin Classifiers, pp. 61–74. MIT-Press (1999)
15. Bichindaritz, I.: Methods in Case-Based Classification in Bioinformatics: Lessons Learned. In: Perner, P. (ed.) ICDM 2011. LNCS, vol. 6870, pp. 300–313. Springer, Heidelberg (2011)
16. Perner, J., Zotenko, E.: Characterizing Cell Types through Differentially Expressed Gene Clusters Using a Model-Based Approach. In: Perner, P. (ed.) ICDM 2011. LNCS, vol. 6870, pp. 106–120. Springer, Heidelberg (2011)