MATHEMATICAL MODELING OF THE DIVERSITY IN HUMAN B AND T CELL RECEPTORS USING MACHINE LEARNING

Susanne Schaller[(a)], Johannes Weinberger[(b)], Martin Danzer[(b)], Christian Gabriel[(b)], Rainer Oberbauer[(c)], Stephan M. Winkler[(a)]

[(a)] Bioinformatics Research Group,
University of Applied Sciences Upper Austria
Softwarepark 13, 4232 Hagenberg, Austria

[(b)] Red Cross Transfusion Service of Upper Austria
Krankenhausstraße 7, 4020 Linz, Austria

[(c)] Elisabethinen Hospital
Fadingerstraße 1, 4020 Linz, Austria

[(a)] susanne.schaller@fh-hagenberg.at, stephan.winkler@fh-hagenberg.at
[(b)] johannes.weinberger@o.roteskreuz.at, martin.danzer@o.roteskreuz.at, christian.gabriel@o.roteskreuz.at
[(c)] rainer.oberbauer@elisabethinen.or.at

ABSTRACT

We here propose an empirical approach based on the analysis of next-generation sequencing (NGS) data for describing the number of distinct clones of B and T-cell receptors in the human immune system. The status of a human immune system is (amongst other features) defined by the diversity of these receptor cells. It is a well-known issue that NGS data have a higher error rate, and therefore the number of distinct sequences found in sequencing data rises with the number of sequences measured by second generation sequencers. We here present a modeling approach that formulates the number of distinct clones depending on the number of read sequences considering two effects. On the one hand there is a true number of distinct sequences which is asymptotically reached by increasing the number of reads, on the other hand the number of randomly found sequences rises linearly due to read errors. The parameters for this combined model are identified using parameter optimization methods using evolution strategies. This modeling approach is evaluated on the basis of immune status data of several human patients. Additionally, the results are compared to those produced by machine learning methods.

Keywords: B and T cell diversity analysis, model identification, parameter identification, immune system, data mining

1. INTRODUCTION: THE HUMAN IMMUNE SYSTEM

The analysis and understanding of the behavior of the human immune system and its key players is one of the most interesting and growing research fields in immunogenetics and medicine. The major cellular components of the adaptive immune response are the B and T cells. These cell types are a subgroup of the white blood cells and are responsible for recognizing foreign antigens and for the immune response to destruct specific pathogens. B cells have the ability to build antibodies which abrogate foreign cells. Surface membrane–bounded antibodies act as B cell receptors on B cells. Similar to the B cells, the T cells play a central role in cell-mediated immunity and can be separated by the T cell receptor, a protein complex which is essential for the recognition of foreign antigens (Elgert, 2009). The function of T cells is to assist other white blood cells in immunologic processes and manage immunological tolerance and memory function. However, T cells are also able to destroy directly virally infected cells and tumor cell.

Modern DNA sequencing systems are referred to as next-generation sequencing (NGS) (Benichou et al., 2012 and Liu et al., 2012). NGS is a method to determine the nucleotide sequence by high-throughput and parallel processing of material, which leads to decreased run times as well as to massively increased amounts of data produced (Tucker et al., 2009). The B and T cell receptor sequence data are measured using next-generation sequencing technologies. Generated sequences are further pre-processed and provide the basis for modeling the diversity of B and T cells in the human immune system to gain more insights into the cell input and into the error rate using these technologies.

2. MATHEMATICAL MODELING OF THE CLONOTYPE DIVERSITY

The research goal of the project presented here is the mathematical modeling of the diversity of B and T cell receptor data. The diversity of B and T cell receptors data provides information about the state of health in humans and supplies detailed insight into the specific existing immune response. This additional know-how helps, e.g., in the diagnostic of autoimmune diseases as well as in transplantation monitoring, and in the treatment of various infections (Boyd et al., 2009).

We perform a modeling approach using data mining and machine learning to identify the relationship to describe the shape and progress of the diversity. So created models shall help to estimate the error rate of the NGS machine and solve additionally biological questions regarding sample preparation and behavior of the NGS technology.

2.1 Biological Data

As first step samples are pre-processed and sequenced using high-throughput NGS technology. B and T cell receptor sequences are further processed using the ImMunoGeneTics (IMGT) system (Alamyar, 2013). The output data derived from the IMGT system are used as input for our diversity analyses (see Figure 1). We are interested in the total number of sequences that are sequenced and the distinct number of clonotypes, where a clonotype represents a unique amino acid sequence.

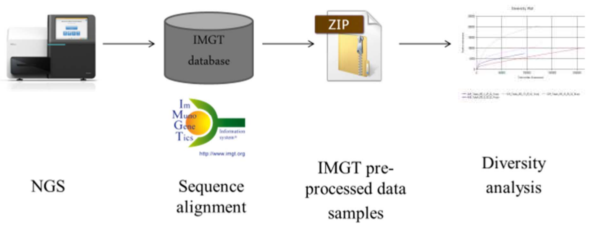

Figure 1: Data sample pre-processing using NGS technology and the IMGT database.

2.2 Diversity Analysis

The diversity analysis plays an important role in the analysis of the human immune system. A mathematical analysis of the diversity of immune cells is of high importance due to two reasons: (1) there is a true number of distinct sequences which can be reached by increasing the number of reads (strings of bases), and (2) the number of randomly found sequences rises linearly due to read errors of the sequencers.

Modelling of the diversity for a biological sample can be constructed by randomly choosing *n* amino acid sequences from the sample and determining the number of distinct sequences, also known as distinct clonotypes of B and/or T cells found in these *n* sequences. This is repeated five times and the number of distinct clones is averaged. The following two examples represent the diversity of B and T cell receptors: Figure 2 represents the diversity of the T cell receptor. The T cell receptor data are gained from blood (4B-TCRB) and kidney (4N-TCRB) samples. Figure 3 represents B cell receptor data.

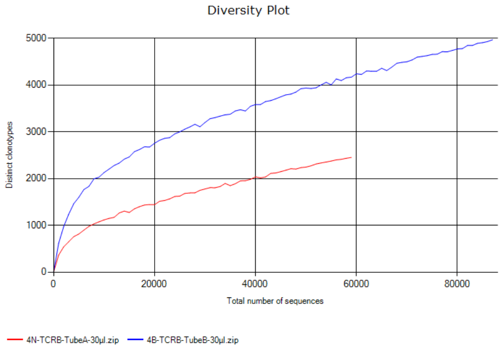

Figure 2: Diversity plot of T cell receptor data (blood and kidney cells).

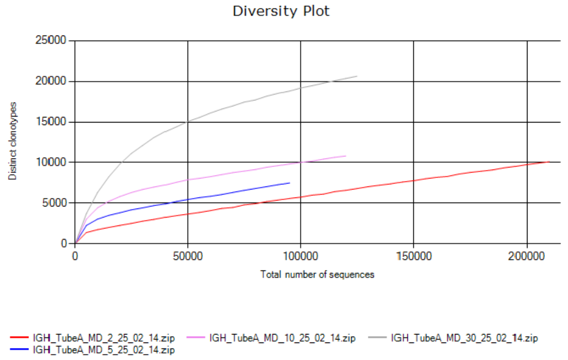

Figure 3: Diversity plot of B cell receptor data (IGH).

3. MODEL IDENTIFICATION

A modeling approach has to be defined that explains the number of distinct sequences found in *n* sequences; this enables us to distinguish between read errors and “real” distinct sequences.

We are especially interested in white box models. In the following sections we present two modeling approaches for which the correct parameters have to be identified empirically for each sample (Section 3.1). Additionally, we also plan to apply symbolic regression using genetic programming for identifying model structures that estimate the number of distinct clonotypes in a biological sample (Section 3.2).

For validating these modeling approaches and for comparing the models’ quality to the modeling quality achievable using alternative approaches, black box models (Section 3.3) shall also be used.

3.1. White Box Modeling Using a Fixed Model Structure and Parameter Optimization

The primary model identification approach used in this work uses different formula structures. The parameters of these models have to be optimized (in this case, using evolution strategies) in order to optimize the fit of the models to the given diversity curves of B and T cell receptor data.

Our modelling possibilities are that the curve initially corresponds to an exponential or logarithmic increase or a multiplicative inverse, while for greater *n* the curve rather resembles a linear function. Therefore we have developed the following model approaches (see Equations 1, 2, 3 and 4) for which the parameters have to be optimized:

$$div(n) = k * n + d \tag{1}$$

$$div(n) = a * (1 - e^{-b*n}) \tag{2}$$

$$div(n) = a * \log(b * n) \tag{3}$$

$$div(n) = a * (1 - e^{-b*n}) + k * n + d \tag{4}$$

For each biological sample, the parameters are optimized using an evolution strategy.

Evolution strategies (ES) have been developed since the 1960s. An ES is an evolutionary algorithm in which genetic operators (mutation, crossover, and selection) are applied on solution candidates until a specific termination criterion is met (Winkler, 2009) and (Borgmann, 2012).

Typically, an ES starts with a population of randomly created individuals; in each generation these individuals are evaluated and new individuals are created using mutation and crossover operators. The next generation's population is formed using successful children. This procedure is repeated over several generations until a termination criterion is met, for example as soon as the quality cannot be improved any more or the maximum number of generations is reached (see Figure 4).

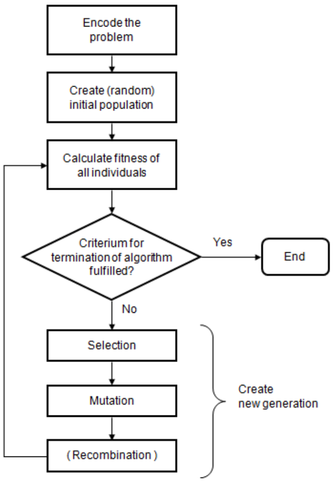
Encode the problem
Create (random) initial population
Calculate fitness of all individuals
Criterium for termination of algorithm fulfilled?
Yes
End
No
Selection
Mutation
(Recombination)
Create new generation

Figure 4: Standard evolution strategy (Winkler, 2009).

One of the goals of this research is to find out, which modeling approach is better in terms of modeling quality, i.e. which approach is more suitable for describing the diversity of B and T cell receptors.

3.2 Black Box Modeling

For estimating the quality of the white box models created using the methods summarized in Sections 3.1, the following black box modeling methods are also applied:

- *Random Forests (RF):* RFs are very popular in the field of machine learning as they are very fast and accurate. They create ensembles of decision trees using information gain and variance reduction. Pruning is done using reduced-error pruning methods (Witten, 2005).

- Genetic Programming: As an alternative to the approach described in the previous section we also apply a modeling algorithm based on genetic programming (Koza, 1992) using a structure identification framework described in (Winkler, 2008 and Affenzeller et al., 2009) implemented in HeuristicLab (Wagner, 2014).

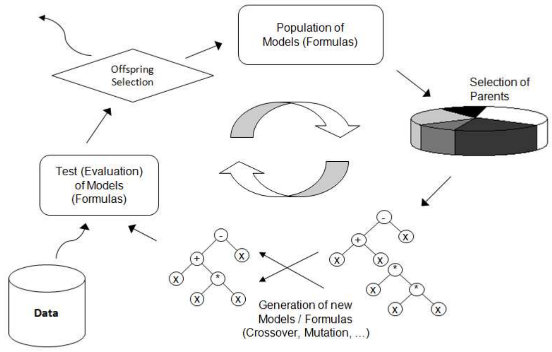
Population of Models (Formulas)
Selection of Parents
Offspring Selection
Test (Evaluation) of Models (Formulas)
Data
Generation of new Models / Formulas (Crossover, Mutation, ...)

Figure 5: Genetic programming approach.

4. RESULTS

4.1 White Box Modeling Experimental Results

For testing the approaches described in Section 2.2 and 3, we used three samples containing a different number of sequences and clonotypes (see Table 1).

	BB17	**SS12**	**6AB**
# sequences	307440	284307	76273
# clonotypes	35906	25899	15580

Table 1: Sequence and clonotype information.

Figures 6, 7, and 8 show the diversity analysis of the following samples BB17, SS12, and 6AB as described in Section 2.2. The thick red line represents the original calculated diversity using iteratively 2500 sequences and determines the number of distinct clonotypes. The red thick lines in Figures 6, 7, and 8 show that there first is a strong increase of the number of distinct clonotypes and then the red line starts to continue at a specific number of sequences to increase more or less linearly. Afterwards where the increase seems to be a more linear slope, which represents the number of randomly found sequences rise linearly due to read errors of the sequencers. All the other thin lines show the estimated diversity curve using four different formulas. The parameters such as *a*, *b*, *k* and *d* are optimized using evolution strategy.

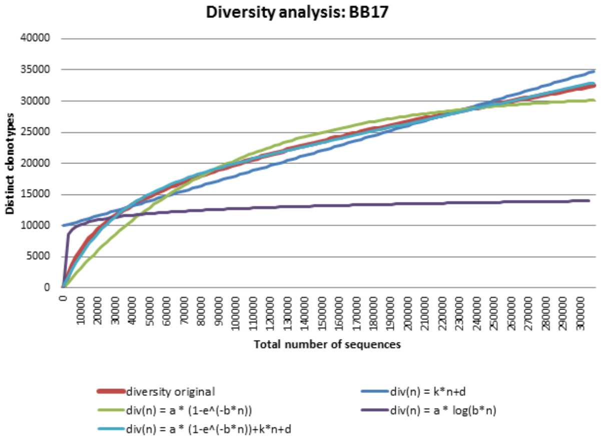

Figure 6: Diversity analysis of sample BB17 using 300 generations, μ=20, λ=100 and σ=0.5.

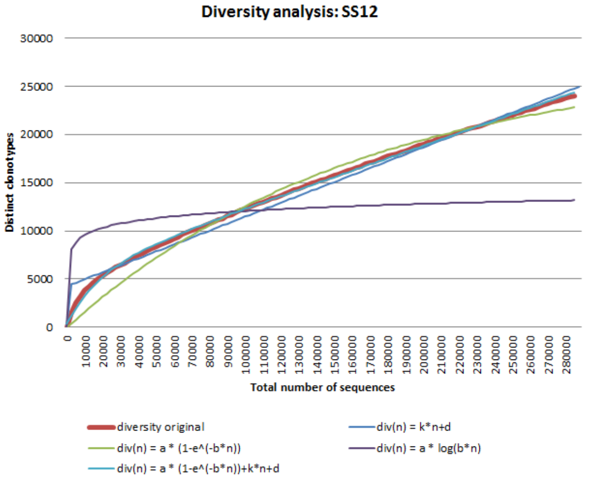

Figure 7: Diversity analysis of sample SS12 using 300 generations, μ=20, λ=100 and σ=0.5.

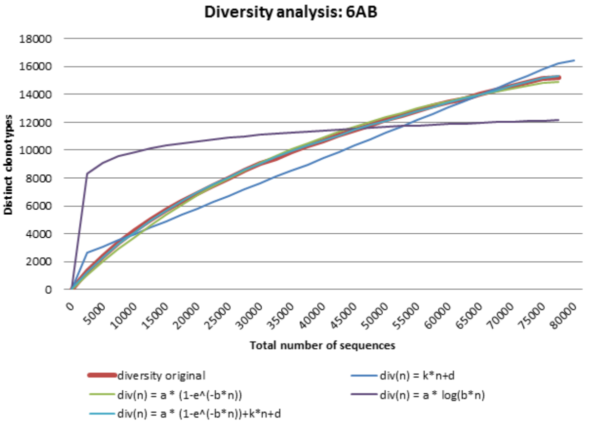

Figure 8: Diversity analysis of sample SS12 using 300 generations, μ=20, λ=100 and σ=0.5.

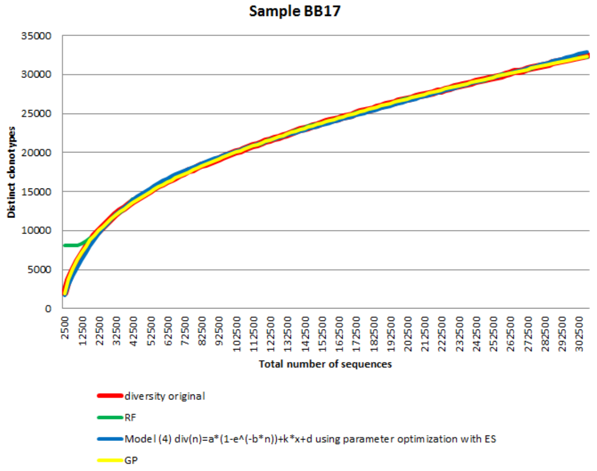

Figure 9: Black box modelling using genetic programming and random forests for sample BB17.

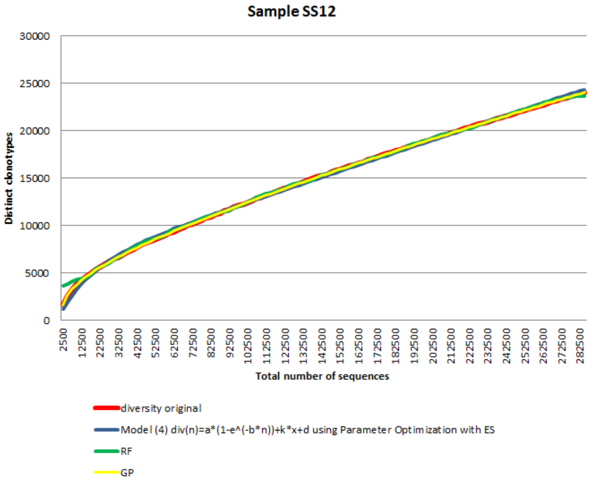

Figure 10: Black box modelling using genetic programming and random forests for sample SS12.

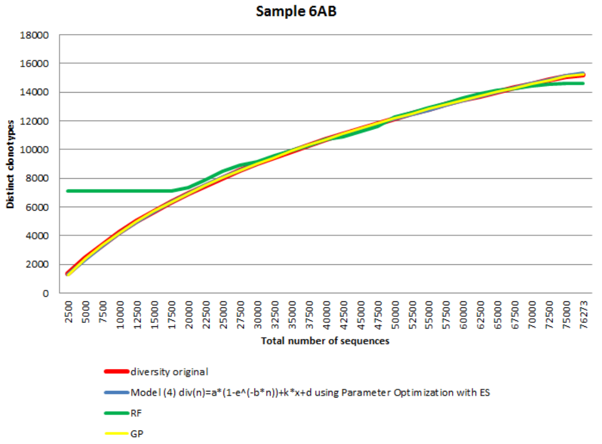

Figure 11: Black box modelling using genetic programming and random forests for sample 6AB.

Here we have used 300 evolution strategy generations and for σ=0.5, μ=20 and λ=100. For the parameter optimization with evolution strategy the so-called 1/5 success rule was used.

The diversity analyses of the samples show that the formulae containing an exponential and a linear function fits the original calculated diversity curve best (light blue line in Figures 6, 7, and 8). Figure 8 demonstrates that in a sample containing a lower total number of sequences also the formulae only using an exponential function (light green line in Figures 6, 7 and 8) perform as well as the combination using a linear function.

To estimate the quality of the white box modelling and to validate the modelling approach we have used the following machine learning methods: genetic programming and random forests.

For black box modelling the framework HeuristicLab (http://dev.heuristiclab.com/) where all those classification and regression algorithms are integrated and implemented. Implementation details of these methods can be found in (Wagner et al., 2014).

For all machine learning algorithms, a different set of parameters have been tested. Best results can be achieved using for genetic programming the following parameter settings: The mutation rate was set to 15% and an OSGA (offspring selection genetic algorithm) (Affenzeller et al., 2009) with success ratio as well as comparison factor set to 1.0 was used. Crossover and mutation were done using the subtree swapping crossover and the multi symbolic expression tree manipulator, respectively. The function set described in (Winkler, 2009) was used for building composite function expressions (including arithmetic, exponential and logarithm functions).

For random forests the following parameter settings have been set: the number of trees has been set to 30 and randomly seed has been used. Generally, data were split into training and test data and five-fold cross-validation was executed. As is shown in Figure 9, 10 and 11 the thick red line represents the original calculated diversity curve, while green and yellow represent the black box modelling machine learning methods. Genetic programming is able to predict the original calculated diversity curve pretty accurate, while using random forest the estimation of the distinct clonotypes seems to be difficult.

Table 2 and 3 represent statistical analysis of the results using the Pearson's R^2 (Bollen et al., 1981) value for training and testing of the white and black box modelling methods. As it is shown that in training and testing very high Pearson's R^2 value can be achieved, where the value 1 means that the originally calculated diversity curve correlates exactly to the predicted diversity curves using machine learning methods such as genetic programming and random forests. In Table 2 it is shown that the highest correlation values can be found using a combination of the exponential and linear function. This high correlation can be observed in all three samples.

Sample/ ES	**div(n)=k*n+d**		**div(n) = a *(1 – $e^{(-b*n)}$)**	
	training	test	training	test
BB17	0,993425	0,992920	0,949096	0,970529
SS12	0,998450	0,998406	0,991846	0,990502
6AB	0,992546	0,953213	0,999550	0,999569

Sample/ ES	**div(n)= a * log(b*n)**		**div(n) = a *(1 – $e^{(-b*n)}$) + k*n+d**	
	training	test	training	test
BB17	0,992889	0,985620	0,998797	0,998740
SS12	0,980015	0,982939	0,999275	0,999172
6AB	0,990095	0,992591	0,999550	0,999569

Table 2: Person's R^2 values for training and testing of all models using parameter optimization with evolution strategy.

Sample/ ML	**GP**		**RF**	
	training	test	training	test
BB17	0,999867	0,999857	0,998261	0,981551
SS12	0,999983	0,999956	0,997973	0,998479
6AB	0,999988	0,999963	0,989975	0,892114

Table 3: Pearson's R^2 values for training and testing of genetic programming and random forests.

5. CONCLUSION

In this paper an approach has been described to model the number of distinct clonotypes in dependence of the total number of sequences (diversity) of B and T receptor cells in the human immune system. It has been shown that there is a true number of distinct clonotypes which can be reached by increasing the number of reads using a mathematical model containing an exponential and linear part. Additionally, we have shown that the diversity analysis curve rises linearly due to read errors of NGS machines. The parameters of the mathematical model are optimized using evolution strategy and for evaluation purposes black box modelling methods such as genetic programming and random forests have been used. The statistical analysis of the results shows that there is a high correlation between the machine learning methods and the parameter optimization approach using evolution strategy.

Currently this analysis is integrated in the framework ImmunExplorer. In future work, additional methods for analyzing B and T cells shall be included in

ImmunExplorer to gain more insight in the behaviour of the human immune system.

ACKNOWLEDGMENTS

The work described in this paper was done within the project “Transplant - Early marker for humoral and cellular rejection of transplanted kidneys” funded by EFRE Regio 13.

REFERENCES

Affenzeller, M., Winkler, S., Wagner, S., Beham, A., 2009. Genetic Algorithms and Genetic Programming - Modern Concepts and Practical Applications. *Chapman & Hall/CRC*.

Alamyar, E., Giudicelli, V., Li, S., Duroux, P., Lefranc, MP., 2012. IMGT/HighV-QUEST: the IMGT(R) web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunome research* 8(1): 26.

Benichou, J., Ben-Hamo, R., Louzoun, Y., Efroni, S. , 2012. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* 135(3):183-91.

Bollen, K.A., Barb, K.H., 1981. Pearson’s R and coarsely categorized measures. *American Sociological Review* – JSTOR.

Borgmann, D., Weghuber, J., Schaller, S., Jacak, J. and Winkler, S.M, 2012. Identification of Patterns in Microscopy Images of Biological Samples using Evolution Strategies. *Proceedings of the 24th European Modeling and Simulation Symposium EMSS*, pp. 271-276.

Boyd, S. D., Marshall, E. L., Merker, J. D., Maniar, J. M., Zhang, L. N., Sahaf, B., Fire, A. Z., 2009. Measurement and Clinical Monitoring of Human Lymphocyte Clonality by Massively Parallel V-D-J Pyrosequencing. *Science Translational Medicine*, 1(12).

Chang, C.C., Lin, C.J., 2001. LIBSVM: a library for support vector machines.

Elgert, K. D., 2009. Immunology: Understanding The Immune System, *John Wiley & Sons*, p. 726.

Koza, JR.., 1994. Genetic programming as a means for programming computers by natural selection. *Statistics and Computing*.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., Law, M., 2012. Comparison of Next-Generation Sequencing Systems. *J Biomed Biotechnol*:251364.

Nelles, O., 2001. Nonlinear System Identification: From Classical Approaches to Neural Networks and Fuzzy Models.

Tucker, T., Marra, M., & Friedman, J. M., 2009. Massively parallel sequencing: the next big thing in genetic medicine. *American Journal of Human Genetics*, 85(2), 142–54.

Vapnik, V., 1998. Statistical Learning Theory. *Wiley, New York*.

Wagner, S., Kronberger, G. K., Beham, A., Kommenda, M., Scheibenpflug, A., Pitzer, E., Vonolfen, S., Kofler, M., Winkler, S. M., Dorfer, V., Affenzeller, M. 2014. Architecture and Design of the HeuristicLab Optimization Environment in Advanced Methods and Applications in Computational Intelligence (Contributions to Book: Part/Chapter/Section 10), (Editors: R. Klempous, J. Nikodem, W. Jacak, Z. Chaczko) - *Springer*, pp. 197-261.

Winkler, S., Affenzeller, M., Wagner, S., 2009. Using enhanced genetic programming techniques for evolving classifiers in the context of medical diagnosis - an empirical study. *Genetic Programming and Evolvable Machines*, pp. 111–140.

Winkler, S., 2009. Evolutionary System Identification - Modern Concepts and Practical Applications. Schriften der Johannes Kepler Universität Linz, Reihe C: Technik und Naturwissenschaften. *Universitätsverlag Rudolf Trauner*.

Witten, I.H., 2005. Data mining: practical machine learning tools and techniques (Second Edition). *Elsevier*.

AUTHORS BIOGRAPHY

SUSANNE SCHALLER received her Masters of Science in Biomedical Informatics in 2009 and 2010 from the University of Skoevde and the University of Applied Sciences Upper Austria. She is a research and teaching associate at the Bioinformatics Research Group, in Hagenberg. Her research interests include systems biology, medical image analysis, tissue engineering, data mining, and machine learning.

JOHANNES WEINBERGER received his Master in Genetics in 2013 from the University of Salzburg. He is currently PhD student at the University of Salzburg and research associate at the Immunogenetics Department of the Red Cross Blood Transfusion Service of Upper Austria in Linz. His research interests are immunology, next-generation sequencing, BCR/TCR spectratyping, and FACS.

MARTIN DANZER received his MSc degree in Genetics from the University of Salzburg. He is actually the Co- Director of the Histocompatibility and Immunogenetics Laboratory in Linz that provides a comprehensive range of diagnostic testing services in the field of transplantation. His research activities focus in particular on immune response gene diversity in the human population and how diversity impacts transplantation.

RAINER OBERBAUER received his MD degree in 1990 at the University of Vienna and received an additional MSc in Epidemiology at the Harvard School of Public Health in 2005. Since 2006 he is director of the Department of Internal Medicine III at the Elisabethinen Hospital in Linz. His research areas focus on nephrology, dialysis and transplantation. Since 2010 Mr. Oberbauer has been teaching medicine at the Medical University of Vienna.

CHRISTIAN GABRIEL studied medicine in Innsbruck from 1978 to 1983. He worked as resident for general medicine from 1984 to 1989, then joined the Department of Anaesthesiology and Intensive Care at the General Hospital in Linz and served as resident for transfusion medicine in the Red Cross Transfusion Service of Upper Austria. Since 2001 he is the medical director of the Red Cross Transfusion Service of Upper Austria in Linz. His research interest includes intensive care, blood physiology, cell therapies, regenerative medicine, and foremost immunogenetics.

STEPHAN M. WINKLER received his PhD in engineering sciences in 2008 from Johannes Kepler University (JKU) Linz, Austria. His research interests include genetic programming, nonlinear model identification and machine learning. Since 2009, Dr. Winkler is professor at the Department for Medical and Bioinformatics at the University of Applied Sciences (UAS) Upper Austria at Hagenberg Campus; since 2010, Dr. Winkler is head of the Bioinformatics Research Group at UAS, Hagenberg.