ON THE USE OF ESTIMATED TUMOR MARKER CLASSIFICATIONS IN TUMOR DIAGNOSIS PREDICTION - A CASE STUDY FOR BREAST CANCER

Stephan M. Winkler [(a)], Michael Affenzeller [(b)], Gabriel Kronberger [(c)], Michael Kommenda [(d)], Stefan Wagner [(e)], Witold Jacak [(f)], Herbert Stekel [(g)]

[(a – f)] Upper Austria University of Applied Sciences
School for Informatics, Communications, and Media
Heuristic and Evolutionary Algorithms Laboratory
Softwarepark 11, 4232 Hagenberg, Austria

[(g)] General Hospital Linz
Central Laboratory
Krankenhausstraße 9, 4021 Linz, Austria

[(a)] stephan.winkler@fh-hagenberg.at, [(b)] michael.affenzeller@fh-hagenberg.at, [(c)] gabriel.kronberger@fh-hagenberg.at, [(d)] michael.kommenda@fh-hagenberg.at, [(e)] stefan.wagner@fh-hagenberg.at, [(f)] witold.jacak@fh-hagenberg.at, [(g)] herbert.stekel@akh.linz.at

ABSTRACT

In this paper we describe the use of tumor marker estimation models in the prediction of tumor diagnoses. In previous work we have identified classification models for tumor markers that can be used for estimating tumor marker values on the basis of standard blood parameters. These virtual tumor markers are now used in combination with standard blood parameters for learning classifiers that are used for predicting tumor diagnoses.

Several data based modeling approaches implemented in HeuristicLab have been applied for identifying estimators for selected tumor markers and cancer diagnoses: Linear regression, k-nearest neighbor learning, artificial neural networks, and support vector machines (all optimized using evolutionary algorithms) as well as genetic programming.

In the results section we summarize classification accuracies for breast cancer; we compare classification results achieved by models that use measured marker values as well as models that use virtual tumor markers.

Keywords: Evolutionary Algorithms, Medical Data Analysis, Tumor Marker Modeling, Data Mining,

1. INTRODUCTION, RESEARCH GOALS

In this paper we present research results achieved within the Josef Ressel Centre for Heuristic Optimization *Heureka!*: Data of thousands of patients of the General Hospital (AKH) Linz, Austria, have been analyzed in order to identify mathematical models for cancer diagnoses. We have used a medical database compiled at the central laboratory of AKH in the years 2005 – 2008: 28 routinely measured blood values of thousands of patients are available as well as several tumor markers (TMs, substances found in humans that can be used as indicators for certain types of cancer). Not all values are measured for all patients, especially tumor marker values are determined and documented only if there are indications for the presence of cancer. The results of empirical research work done on the data based identification of estimation models for cancer diagnoses are presented in this paper: Based on patients' data records including standard blood parameters, tumor markers, and information about the diagnosis of tumors we have trained mathematical models for estimating tumor markers and cancer diagnoses.

In previous work (Winkler et al. 2010; Winkler et al. 2011) we have identified classification models for tumor markers that can be used for estimating tumor marker values on the basis of standard blood parameters. These tumor marker models (also referred to as virtual markers) are now used in combination with standard blood parameters for learning classifiers that can be used for predicting tumor diagnoses. Our goal is to show to which extent virtual tumor markers can replace tumor marker measurements in the prediction of cancer diagnoses.

These research goals and the respective modeling tasks are graphically summarized in Figure 1.

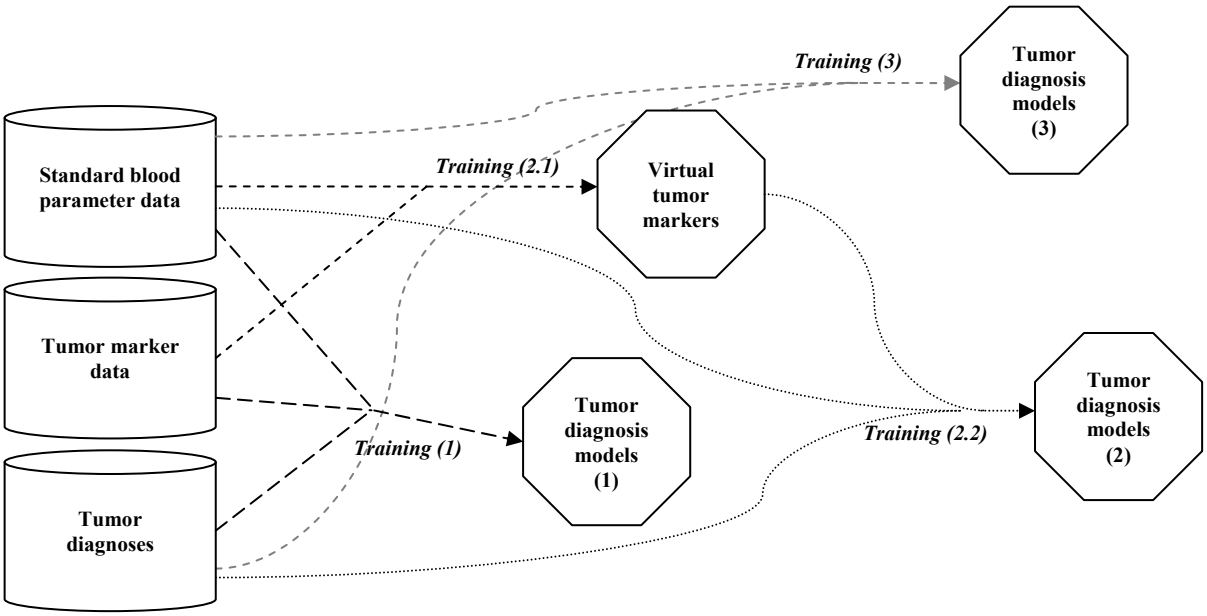

Figure 1: Modeling tasks addressed in this research work: Tumor markers are modeled using standard blood parameters and tumor markers (training scenario 1); tumor diagnosis models are trained using standard blood values and on the one hand tumor marker data and on the other hand using estimated tumor markers (scenario 2); alternatively we also train diagnosis estimation models only using standard blood parameters and diagnosis information (scenario 3).

- First, models are trained for estimating tumor diagnoses using the full set of available data: Standard blood parameters, tumor marker data and diagnoses information are used.
- Second, machine learning is applied for learning estimation models for tumor markers. Concretely, we identify classification models that predict tumor marker values for given standard blood parameters either as "normal" or as "elevated". Subsequently models are trained for estimating tumor diagnoses using standard blood parameters and diagnoses data, but instead of tumor marker values the estimated tumor marker classifications (calculated using models learned in the first modeling step) are used.
- Third, we also train diagnosis estimation models only using standard blood parameters and diagnosis information.

In the following section (Section 2) we summarize the machine learning methods that were applied in this research project, in Section 3 we give an overview of the data base that was used, and in Section 4 we document the modeling results that could be achieved. This paper is concluded in Section 5.

2. MACHINE LEARNING METHODS APPLIED

In this section we describe the modeling methods applied for identifying estimation models for tumor markers and cancer diagnoses: On the one hand we apply hybrid modeling using machine learning algorithms and evolutionary algorithms for parameter optimization and feature selection (as described in Section 2.1), on the other hand we use genetic programming (as described in Section 2.2). In (Winkler et al. 2011), for example, these methods have also been described in detail.

2.1. Hybrid Modeling Using Machine Learning Algorithms and Evolutionary Algorithms for Parameter Optimization and Features Selection

Feature selection is often considered an essential step in data based modeling; it is used to reduce the dimensionality of the datasets and often conducts to better analyses. Given a set of n features $F = \{f_1, f_2, \ldots, f_n\}$, our goal here is to find a subset of F, F', that is on the one hand as small as possible and on the other hand allows modeling methods to identify models that estimate given target values as well as possible. Additionally, each data based modeling method (except plain linear regression) has several parameters that have to be set before starting the modeling process.

The fitness of feature selection F' and training parameters with respect to the chosen modeling method is calculated in the following way: We use a machine learning algorithm m (with parameters p) for estimating predicted target values $est(F',m,p)$ and compare those to the original target values $orig$; the coefficient of determination R^2 function is used for calculating the quality of the estimated values. Additionally, we also calculate the ratio of selected features $|F'|/|F|$. Finally, using a weighting factor α, we calculate the fitness of the set of features F' using m and p as

$$fitness(F', m, p) = \alpha \cdot |F'|/|F| + (1-\alpha) \cdot (1-R^2(est(F',m,p),orig)). \quad (1)$$

In (Alba et al. 2007), for example, the use of evolutionary algorithms for feature selection optimization is discussed in detail in the context of gene selection in cancer classification. We have now used

evolutionary algorithms for finding optimal feature sets as well as optimal modeling parameters for models for tumor diagnosis; this approach is schematically shown in Figure 2: A solution candidate is here represented as $[s_{1...n}p_{1...q}]$ where s_i is a bit denoting whether feature F_i is selected or not and p_j is the value for parameter j of the chosen modeling method m. This rather simple definition of solution candidates enables the use of standard concepts for genetic operators for crossover and mutation of bit vectors and real valued vectors: We use uniform, single point, and 2-point crossover operators for binary vectors and bit flip mutation that flips each of the given bits with a given probability. Explanations of these operators can for example be found in (Holland 1975) and (Eiben 2003).

We have used strict offspring selection (Affenzeller et al. 2009): Individuals are accepted to become members of the next generation if they are evaluated better than both parents.

In (Winkler et al. 2011) we have documented classification accuracies for tumor diagnoses using this approach for optimizing feature set and modeling parameters.

The following machine learning algorithms have been applied for identifying estimators for selected tumor markers and cancer diagnoses: Linear regression, k-nearest neighbor learning, artificial neural networks, and support vector machines.

2.2. Genetic Programming

As an alternative to the approach described in the previous sections we have also applied a classification algorithm based on genetic programming (GP, Koza (1992)) using a structure identification framework described in Winkler (2008) and Affenzeller et al. (2009).

We have used the following parameter settings for our GP test series: The mutation rate was set to 20%, gender specific parents selection (Wagner 2005) (combining random and roulette selection) was applied as well as strict offspring selection (Affenzeller et al. 2009) (OS, with success ratio as well as comparison factor set to 1.0). The functions set described in (Winkler 2008) (including arithmetic as well as logical ones) was used for building composite function expressions.

In addition to splitting the given data into training and test data, the GP based training algorithm used in our research project has been designed in such a way that a part of the given training data is not used for training models and serves as validation set; in the end, when it comes to returning classifiers, the algorithm returns those models that perform best on validation data.

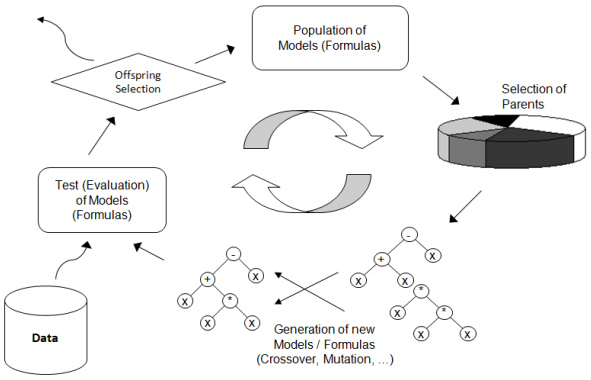

Figure 2: Genetic programming including offspring selection.

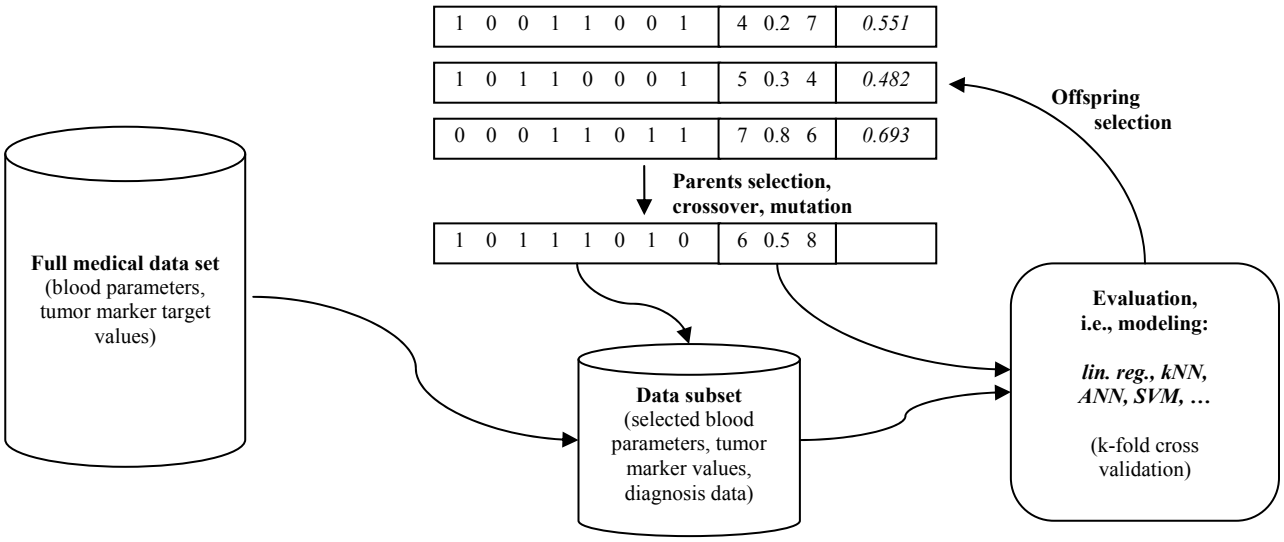

Figure 3: A hybrid evolutionary algorithm for feature selection and parameter optimization in data based modeling. Machine learning algorithms are applied for evaluating feature sets.

3. DATA BASIS

3.1. General Information

The blood data measured at the AKH in the years 2005-2008 have been compiled in a database storing each set of measurements (belonging to one patient): Each sample in this database contains an unique ID number of the respective patient, the date of the measurement series, the ID number of the measurement, standard blood parameters, tumor marker values, and cancer diagnosis information. Patients' personal data were at no time available for the authors except for the head of the laboratory.

In total, information about 20,819 patients is stored in 48,580 samples. Please note that of course not all values are available in all samples; there are many missing values simply because not all blood values are measured during each examination. Further details about the data set and necessary data preprocessing steps can for example be found in Winkler et al. (2010) and Winkler et al. (2011), e.g.

Standard blood parameters include for example the patients' sex and age, information about the amount of cholesterol and iron found in the blood, the amount of hemoglobin, and the amount of red and white blood cells; in total, 29 routinely available patient parameters are available.

Literature discussing tumor markers, their identification, their use, and the application of data mining methods for describing the relationship between markers and the diagnosis of certain cancer types can be found for example in Koepke (1992) (where an overview of clinical laboratory tests is given and different kinds of such test application scenarios as well as the reason of their production are described) and Yonemori (2006).

Information about the following tumor markers is stored in the AKH database: AFP, CA 125, CA 15-3, CA 19-9, CEA, CYFRA, fPSA, NSE, PSA, S-100, SCC, and TPS.

Finally, information about cancer diagnoses is also available in the AKH database: If a patient is diagnosed with any kind of cancer, then this is also stored in the database. Our goal in the research work described in this paper is to identify estimation models for the presence of the breast cancer, cancer class C50 according to the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10).

3.2. Data Preprocessing

Before starting the modeling algorithms for training classifiers we had to compile a data set specific for breast cancer diagnosis:

- First, blood parameter measurements were joined with diagnosis results; only measurements and diagnoses with a time delta less than a month were considered.
- Second, all samples containing measured values for breast cancer are extracted.
- Third, all samples are removed that contain less than 15 valid values.
- Finally, variables with less than 10% valid values are removed from the data base.

This leads to a data set specific for breast cancer; this *BC* data set consists of 706 samples with 324 samples (45.89%) labeled with '0' and 382 (54.11%) labeled with '1'.

The following variables are stored in this so compiled data set *BC*: Age, sex, tumor diagnosis (0/1), ALT, AST, BSG1, BUN, C125 (TM), C153 (TM), , CBAA, CEA (TM), CEOA, CH37, CHOL, CLYA, CMOA, CNEA, CRP, FE, FER, GT37, HB, HDL, HKT, HS, KREA, LD37, MCV, PLT, RBC, TBIL, TF, and WBC. Three tumor markers (C125, C153, and CEA) are available in *BC*.

4. MODELING TEST SERIES

We have trained models for estimating breast cancer diagnoses using genetic programming as described in Section 2.2 with strict OS and population size 200; as rather small and compact models are preferred by the authors' medical cooperation partners, the maximum size of the evolved models was set to 20, 35, and 50 nodes. For modeling and test results achieved using bigger model structures please see Winkler et al. (2011).

For training virtual tumor markers we have used the hybrid modeling approach described in Section 2.1: 5-fold cross validation was applied, linear regression (linReg), k-nearest-neighbor (kNN) learning, artificial neural networks (ANNs), and support vector machines (SVMs) have been used as machine learning algorithms, and their feature selections and modeling parameters have been optimized by an evolutionary algorithm. Details about these machine learning approaches and their implementation can for example be found in Winkler et al. (2010) and Winkler et al. (2011).

We have used the implementations in the open source framework HeuristicLab (Wagner (2009)) (http://dev.heuristiclab.com).

4.1. Modeling Strategies

For training estimation models for breast cancer we have applied four different strategies: One using measured tumor markers, one using virtual tumor marker classifiers (combined with OR-conjunctions), one using virtual tumor marker classifiers (combined with majority voting), and finally one not using tumor markers at all.

4.1.1. Strategy I: Using standard blood parameters and measured tumor markers

Measured tumor markers were used as well as standard blood parameters, classification models for breast cancer diagnoses were trained using GP as described above.

This corresponds to scenario 1 as described in Section 1.

4.1.2. Strategy II: Using standard blood parameters and virtual tumor marker classifiers

We have used hybrid modeling as described in Section 2.1 for creating virtual tumor marker classifiers on the basis of the data available in the BC data set. The population size for the optimization algorithm (a GA with strict OS) was set to 20, the maximum selection pressure to 100, and α to 0.1; the test classifications of the so identified best virtual tumor marker classifiers were used as estimated tumor marker values. Classification models for breast cancer diagnoses were trained using GP as described above using virtual tumor markers as well as standard blood parameters.

This corresponds to scenario 2 as described in Section 1.

We applied each machine learning algorithm used here (namely linear regression, support vector machines, neural networks, and kNN classification) twice in each modeling process, leading to eight estimated binary classifications for each tumor marker; these classifications were combined into one binary classification variable for each tumor marker using either an OR conjunction or majority voting:

Strategy II.a: Using OR: If any of the classifiers for a tumor marker return 1 for a sample, then this sample's virtual tumor marker is 1; else it is set to 0.

Strategy II.b: Using majority voting: If more than the half of the classifiers for a tumor marker return 1 for a sample, then this sample's virtual tumor marker is 1; else it is set to 0.

4.1.3. Strategy III: Using only standard blood parameters

In this strategy no tumor markers were used; instead, only standard blood parameters were available for training classification models for breast cancer diagnoses using GP as described above.

This corresponds to scenario 3 as described in Section 1.

4.2. Test Results

As already mentioned, each strategy was executed using five-fold cross validation; we here report on the average classification accuracies on test samples ($\mu \pm \sigma$) which are also shown in Figure 4:

- Strategy I: 0.777 ± 0.104
- Strategy II.a: 0.713 ± 0.107
- Strategy II.b: 0.752 ± 0.042
- Strategy III: 0.699 ± 0.113

5. CONCLUSIONS

In the test results summarized in Section 4 we see that the virtual tumor markers have turned out to be able to improve classification accuracy for the modeling application described in this paper: Whereas classifiers not using tumor markers classify approximately 70% of the samples correctly, the use of virtual tumor markers (combined using majority voting) leads to an increase of the classification accuracy to ~75%. Still, virtual tumor markers have in this example not been able to replace measured tumor markers perfectly, as the classification accuracy of models using measured TMs reaches ~77.7%.

Future work on this topic shall include the investigation of virtual TMs for other types of diseases; furthermore, we will also focus on the practical application of the here presented research results in the treatment of patients.

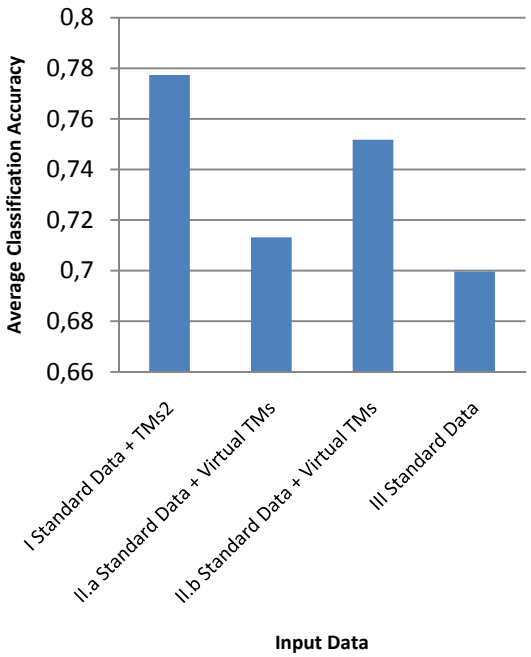

Figure 4: Average test accuracies achieved using the four modeling strategies described in Section 4.1

ACKNOWLEDGMENTS

The work described in this paper was done within the Josef Ressel Centre for Heuristic Optimization *Heureka!* (http://heureka.heuristiclab.com/) sponsored by the Austrian Research Promotion Agency (FFG).

REFERENCES

Affenzeller, M., Winkler, S., Wagner, S., A. Beham, 2009. *Genetic Algorithms and Genetic Programming - Modern Concepts and Practical Applications*. Chapman & Hall/CRC. ISBN 978-1584886297. 2009.

Alba, E., García-Nieto, J., Jourdan, L., Talbi, E.-G., 2005. Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. *IEEE Congress on Evolutionary Computation 2007*, pp. 284 – 290.

Eiben, A.E. and Smith, J.E. 2003. Introduction to Evolutionary Computation. *Natural Computing Series*, Springer-Verlag Berlin Heidelberg.

Holland, J.H., 1975. *Adaption in Natural and Artifical Systems*. University of Michigan Press.

Koepke, J.A., 1992. Molecular marker test standardization. *Cancer*, 69, pp. 1578–1581.

Rechenberg, I., 1973. *Evolutionsstrategie.* Friedrich Frommann Verlag.

Schwefel, H.-P., 1994. *Numerische Optimierung von Computer-Modellen mittels der Evolutions-strategie.* Basel: Birkhäuser Verlag.

Wagner, S., 2009. *Heuristic Optimization Software Systems - Modeling of Heuristic Optimization Algorithms in the HeuristicLab Software Environment*. PhD Thesis, Institute for Formal Models and Verification, Johannes Kepler University Linz, Austria.

Winkler, S., Affenzeller, M., Jacak, W., Stekel, H., 2010. Classification of Tumor Marker Values Using Heuristic Data Mining Methods. *Proceedings of Genetic and Evolutionary Computation Conference 2010, Workshop on Medical Applications of Genetic and Evolutionary Computation*, pp. 1915–1922.

Winkler, S., Affenzeller, M., Jacak, W., Stekel, H., 2011. Identification of Cancer Diagnosis Estimation Models Using Evolutionary Algorithms – A Case Study for Breast Cancer, Melanoma, and Cancer in the Respiratory System. *Proceedings of Genetic and Evolutionary Computation Conference 2011, Workshop on Medical Applications of Genetic and Evolutionary Computation*.

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. ISBN 978-3-85499-569-2.

Yonemori, K., Ando, M., Taro, T. S., Katsumata, N., Matsumoto, K., Yamanaka, Y., Kouno, T., Shimizu, C., Fujiwara, Y., 2006. Tumor-marker analysis and verification of prognostic models in patients with cancer of unknown primary, receiving platinum-based combination chemotherapy. *Journal of Cancer Research and Clinical Oncology*, 132(10), pp. 635–642.

AUTHORS BIOGRAPHIES

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.

MICHAEL AFFENZELLER has published several papers, journal articles and books dealing with theoretical and practical aspects of evolutionary computation, genetic algorithms, and meta-heuristics in general. In 2001 he received his PhD in engineering sciences and in 2004 he received his habilitation in applied systems engineering, both from the Johannes Kepler University of Linz, Austria. Michael Affenzeller is professor at UAS, Campus Hagenberg, and head of the Josef Ressel Center *Heureka!* at Hagenberg.

GABRIEL KRONBERGER received his PhD in engineering sciences in 2010 from JKU Linz, Austria, and is a research associate at the UAS Research Center Hagenberg. His research interests include genetic programming, machine learning, and data mining and knowledge discovery.

MICHAEL KOMMENDA finished his studies in bioinformatics at Upper Austria University of Applied Sciences in 2007. Currently he is a research associate at the UAS Research Center Hagenberg working on data-based modeling algorithms for complex systems within *Heureka!*.

STEFAN WAGNER received his PhD in engineering sciences in 2009 from JKU Linz, Austria; he is professor at the Upper Austrian University of Applied Sciences (Campus Hagenberg). Dr. Wagner's research interests include evolutionary computation and heuristic optimization, theory and application of genetic algorithms, and software development.

WITOLD JACAK received his PhD in electric engineering in 1977 from the Technical University Wroclaw, Poland, where he was appointed Professor for Intelligent Systems in 1990. Since 1994 Prof. Jacak is head of the Department for Software Engineering at the Upper Austrian University of Applied Sciences (Campus Hagenberg) where he currently also serves as Dean of the School of Informatics, Communications and Media.

HERBERT STEKEL received his MD from the University of Vienna in 1985. Since 1997 Dr. Stekel is chief physician at the General Hospital Linz, Austria, where Dr. Stekel serves as head of the central laboratory.