# FEATURE SELECTION IN THE ANALYSIS OF TUMOR MARKER DATA USING EVOLUTIONARY ALGORITHMS

**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 evolutionary algorithms for the selection of relevant features in the context of tumor marker modeling. Our aim is to identify mathematical models for classifying tumor marker values AFP and CA 15-3 using available patient parameters; data provided by the General Hospital Linz are used. The use of evolutionary algorithms for finding optimal sets of variables is discussed; we also define fitness functions that can be used for evaluating feature sets taking into account the number of selected features as well as the resulting classification accuracies.

In the empirical section of this paper we document results achieved using an evolution strategy in combination with several machine learning algorithms (linear regression, k-nearest-neighbor modeling, and artificial neural networks) which are applied using cross-validation for evaluating sets of selected features. The identified sets of relevant variables as well as achieved classification rates are compared.

Keywords: Evolutionary Algorithms, Medical Data Analysis, Tumor Marker Modeling, Data Mining, Machine Learning, Classification, Statistical Analysis

## 1. INTRODUCTION AND SCIENTIFIC GOALS

In general, tumor markers are substances found in humans (especially blood and / or body tissues) that can be used as indicators for certain types of cancer. There are several different tumor markers which are used in oncology to help detect the presence of cancer; elevated tumor marker values can indicate the presence of cancer, but there can also be other causes. As a matter of fact, elevated tumor marker values themselves are not diagnostic, but rather only suggestive; tumor markers can be used to monitor the result of a treatment (as for example chemotherapy). 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 et al., 2006).

The general goal of the research work described here is to identify models for estimating selected tumor marker values on the basis of routinely available blood values; in detail, estimators for the tumor markers AFP and CA 15-3 have been identified. The documented tumor marker values are classified as “normal”, “slightly elevated”, “highly elevated”, and “beyond plausible”; our goal is to design classifiers for the 2-class-classification problem classifying samples into “normal” vs. “elevated”, “highly elevated”, or “beyond plausible”.

In the research work reported on in this paper we use evolutionary algorithms for optimizing the selection of features that are used by machine learning algorithms for modeling the given target values. This approach is closely related to the method described in (Alba, García-Nieto, Jourdan, and Talbi 2007) where the authors compared the use of a particle swarm optimization (PSO) and a genetic algorithm (GA), both augmented with support vector machines, for the classification of high dimensional microarray data.

## 2. DATA BASE

Data of thousands of patients of the General Hospital (AKH) Linz, Austria, have been analyzed in order to identify mathematical models for tumor markers. We have used a medical data base compiled at the Central

Laboratory of the General Hospital Linz, Austria, in the years 2005 – 2008: 28 routinely measured blood values of thousands of patients are available as well as several tumor markers; not all values are measured for all patients, especially tumor marker values are determined and documented if there are indications for the presence of cancer.

The blood data measured at the AKH in the years 2005-2008 have been compiled in a data base storing each set of measurements (belonging to one patient): Each sample in this data base contains an unique ID number of the respective patient, the date of the measurement series, the ID number of the measurement, and several other clinical parameters; standard blood parameters are stored as well as tumor marker values. Data that could identify patients uniquely (as for example name, date of birth, …) where at no time available to the authors except 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.

In (Winkler, Affenzeller, Jacak, and Stekel 2010) the authors give further details about the data used in the research work described here; background information about the blood parameters given in the following two subsections as well as references to important literature on these clinical parameters can be found there, too.

### 2.1. Input Data

The following features are available in the AKH data base and are potential inputs for modeling the given tumor marker values: ALT (lanine transaminase), AST (aspartate transaminase), BSG1 (the erythrocyte sedimentation rate), BUN (blood urea nitrogen), CBAA (basophil granulocytes), CEOA (eosinophil granulocytes), CH37 (cholinesterase), CHOL (cholesterol), CLYA (lymphocytes), CMOA (monocytes), CNEA (neutrophils), CRP (c-reactive protein), FE (iron), FER (ferritin), GT37 (γ-glutamyltransferase), HB (hemoglobin), HDL (high-density lipoprotein), HKT (hematocrit), HS (uric acid), KREA (creatinine), LD37 (lactate dehydrogenase) MCV (mean corpuscular / cell volume), PLT (thrombocytes), RBC (erythrocytes), TBIL (bilirubin), TF (transferring), WBC (leukocytes), and finally the age and the sex of the patients.

### 2.2. Target Data

In addition to the features listed in the previous section, the following tumor markers are also documented in the AKH data base: AFP (alpha-fetoprotein), CA 125 (cancer antigen 125), CA 15-3 (mucin 1), CEA (carcinoembryonic antigen), CYFRA (fragments of cytokeratin 19), and PSA (prostate-specific antigen).

The two tumor markers analyzed in this work can be described in the following way:

- ***AFP:*** Alpha-fetoprotein is a protein found in the blood plasma; during fetal life it is produced by the yolk sac and the liver. For example, AFP values of pregnant women can be used in screening tests for developmental abnormalities as increased values might for example indicate open neural tube defects, decreased values might indicate Down syndrome. AFP is also often measured and used as a marker for a set of tumors, especially endodermal sinus tumors (yolk sac carcinoma), neuroblastoma, hepatocellular carcinoma, and germ cell tumors (Duffy and Crown, 2008).
- ***CA 15-3:*** Mucin 1 (MUC1), also known as cancer antigen 15-3 (CA 15-3), is a protein found in humans; it is used as a tumor marker in the context of monitoring certain cancers (Niv, 2008), especially breast cancer.

### 2.3. Data Preprocessing

Before analyzing the data and using them for training classifiers we have preprocessed the available data:

First, all variables have been linearly scaled to the interval [0;1]: For each variable $v_i$, the (predefined) minimum value $min_i$ is subtracted from all contained values and the result divided by the difference between $min_i$ and the (also predefined) maximum plausible value $maxplau_i$; all values greater than the given maximum plausible value are replaced by 1.0. Then, all samples belonging to the same patient with not more than one day difference with respect to the measurement data have been merged. This has been done in order to decrease the number of missing values in the data matrix. In rare cases, more than one value might thus be available for a certain variable; in such a case, the first value is used.

Additionally, all measurements have been sample-wise re-arranged and clustered according to the patients' IDs; this has been done in order to prevent data of certain patients being included in the training as well as in the test data.

Before using modeling algorithms for training classifiers we have compiled separate data sets for each analyzed target tumor marker $tm_i$: First, all samples containing measured values for $tm_i$ are extracted. Second, all variables are removed from the resulting data set that contain values in less than 80% of the remaining samples. Third, all samples are removed that still contain missing values. This procedure results in a specialized data set $dstm_i$ for each tumor marker $tm_i$.

Details about these data preprocessing steps can be found in (Winkler, Affenzeller, Jacak, and Stekel 2010).

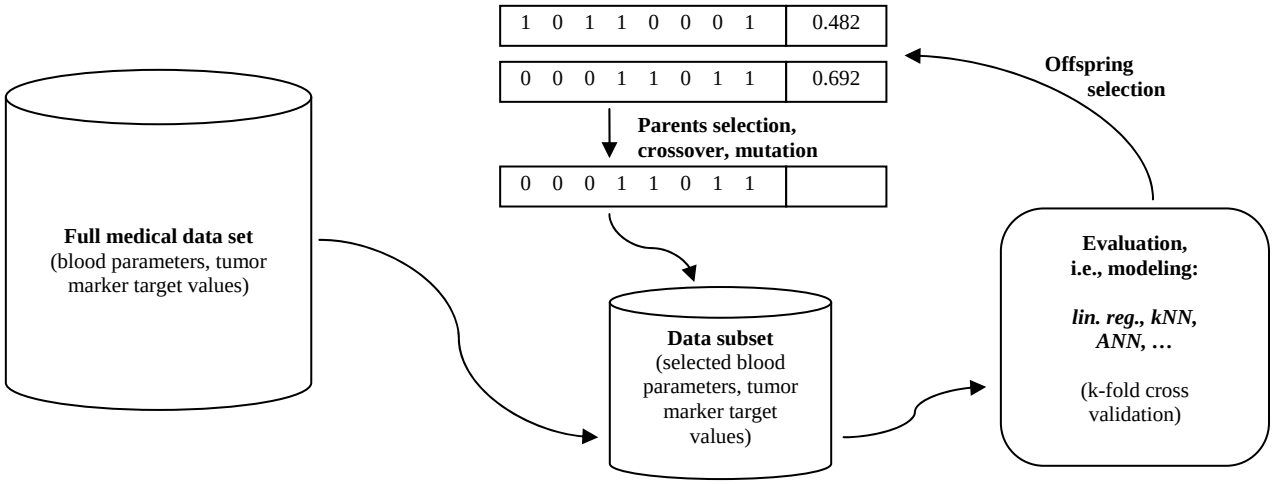


Figure 1: A hybrid evolutionary algorithm for feature selection in the context of tumor marker modeling. Machine learning algorithms are applied for evaluating feature sets.

## 3. MODELING APPROACH

As for example discussed in detail in (Alba, García-Nieto, Jourdan, and Talbi 2007), feature selection is often considered an essential step in data analysis as this method can reduce the dimensionality of the datasets and often conducts to better analyses.

Given a set of features $F = \{f_1, f_2, \ldots, f_n\}$, our goal here is to find a subset $F' \subseteq F$ 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. This implies that we have to deal with a tradeoff situation as in most cases we will see that more complex models (i.e., models using more features) show better fit on data.

This optimization approach is schematically shown in Figure 1.

The fitness of a feature selection $F_k$ is calculated in the following way: We use a machine learning method (linear regression, k-nearest-neighbor modeling, artificial neural networks, SVMs, or any other kind of training algorithm) for estimating predicted target values $est_k$ 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, i.e., we calculate the quality of the model produced using $F_k$, $r^2(F_k)$: $r^2(F_k) = r^2(est_k,orig)$. Additionally, we also calculate the ratio of selected features $a$ as $a(F_k,F) = ||F_k||/||F||$. Finally, using a weighting factor $\alpha$, we calculate the fitness of the set of features $F_k$:

$$fitness(F_k) = \alpha * a(F_k,F) + (1- \alpha) * (1-r^2(F_k)) \quad (1)$$

As an alternative to the coefficient of determination function we can also use a classification specific function that calculates the ratio of correctly classified samples, either in total or as the average of all classification accuracies of the given classes (as for example described in (Winkler 2009), Section 8.2):

For all samples that are to be considered we know the original classifications *origClass*, and using (predefined or dynamically chosen) thresholds we get estimated classifications $estClass_k$ for estimated target values $est_k$ (calculated using features set $F_k$). The total classification accuracy $ca_k$ is calculated as

$$ca(F_k) = ||\{j: estClass_k[j] = origClass[j]\}|| / ||estClass|| \quad (2)$$

Class-wise classification accuracies *cwca* are calculated as the average of all classification accuracies for each given class $c \in C$, $ca_{k,c}$, separately:

$$ca_{k,c} = ||\{j: estClass_k[j] = origClass[j] \; \& \; origClass[j] = c\}|| / ||j: origClass[j] = c|| \quad (3)$$

$$cwca(F_k) = \Sigma_{c \in C}(ca_{k,c}) / ||C|| \quad (4)$$

We can now define the classification specific fitness of feature selection Fk as

$$fitness_{class}(F_k) = \alpha * a(F_k,F) + (1- \alpha) * (1-ca(F_k)) \quad (5)$$

or

$$fitness_{class}(F_k) = \alpha * a(F_k,F) + (1- \alpha) * (1-cwca(F_k)) \quad (6)$$

We use evolutionary algorithms for search for solutions that minimize the given fitness functions. Evolutionary algorithms are used for generating solutions consisting of bit vectors $b_j$, where $b_j(m)$ denotes whether variable $m$ is used by solution candidate $j$ or not. This rather simple definition of solution candidates enables the use of standard genetic operators for crossover and mutation of bit 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 and Smith, 2003).

We have implemented and tested several EAs for guiding the search for optimal feature sets, especially evolution strategies (ES) (Rechenberg 1973; Schwefel 1994) and genetic algorithms (GAs) (Holland 1975).

When applying the ES based approach, populations consisting of $\mu$ individuals are used; in each generation, $\lambda$ children are generated using random parent selection, recombination (which is optional) and mutation, and then the $\mu$ best solutions (selected from

the children, if the *comma* strategy is chosen, or from parents and children, if the *plus* strategy is chosen) become the next generation's members. Additionally, the ratio of successful mutations (i.e., mutations that lead to an improvement of the solutions' qualities) is monitored; if this ratio is less than 1/5, then the mutation operator's flip probability is (in our implementation) multiplied by 0.9, and if it is greater than 1/5, then it is divided by 0.9.

The GA approaches used in the authors' research work include the application of parents selection (proportional or linear rank, e.g.), crossover, and mutation. Optionally, offspring selection can also be applied: After generating new solutions by crossover and mutation, these new solutions are inserted into the next generation's population only if they are better than their parents. Details about this procedure can for example be found in (Affenzeller, Winkler, Wagner, and Beham 2009).

## 4. IMPLEMENTATION

The approach described in this paper (including the evolutionary algorithms used for optimizing feature selection as well as the machine learning methods used for evaluating feature sets) have been implemented using the HeuristicLab (HL) framework (http://www.heuristiclab.com; (Wagner 2009)), a framework for prototyping and analyzing optimization techniques for which both generic concepts of evolutionary algorithms and many functions to evaluate and analyze them are available; we have used these implementations for producing the results summarized in the following section.

## 5. EMPIRICAL TESTS AND RESULTS

### 5.1. Test Series Setup

In this section we document modeling results for tumor markers AFP and CA 15-3. We have applied the evolutionary modeling and feature selection approach described in Section 3 using an (10+100) evolution strategy (without recombination, only using mutation); the number of evaluations (of the ES algorithm) was for some test runs limited to 1,000 and for others to 10,000, and the parameter $\alpha$ (which weights the ratio of selected features, see Section 3) was in some cases set to 0.2 and in other to 0.5. Initially, each feature was selected by each solution candidate with 20% probability. When mutating a solution candidate, the initial bit flip probability (i.e., the probability of changing the information about using a certain feature) for each feature was set to 0.5, and this bit flip probability was (as described in Section 3) modified with respect to success ratio; the minimum value for the bit flip probability was set to 0.1.

We have tested each algorithm setting 5 times using different machine learning algorithms (with varying algorithm specific parameter settings); 5-fold cross validation was applied. Linear regression (linReg), k-nearest-neighbor (kNN) learning, and artificial neural networks (ANNs) have been used as machine learning algorithms; details about their implementation in HL can for example be found in (Winkler, Affenzeller, Jacak, and Stekel 2010). The number of neighbors (*k*) considered in kNN classification was varied and set to 3, 5, and 10, respectively; for training ANNs the number of hidden nodes was set to 10, and 30% of the available training samples were used as validation set, i.e., the network that performed best on validation samples was eventually presented as result of the learning process.

For evaluating modeling results, the fitness function (1) given in Section 3 was used. In the following subsections we summarize the results documented for these test parameters; for each tumor marker and each modeling scenario we list the input features selected by the evolutionary process (features not selected in each test run are given in brackets) as well as average best fitness values (which are used by the evolutionary algorithm and are calculated using validation samples) and test classification accuracies (i.e., the ratios of correctly classified samples that were not considered during the learning process).

### 5.2. Test Results for AFP

Table 1: Results for AFP (linReg; $\alpha = 0.2$)

| Evaluations | Selected features | Best fitness | Avg. test accuracy |
|---|---|---|---|
| 1,000 | (AGE), AST, (CH37), (HKT), (KREA), (PLT), | 0.6867 | 80.73% |
| 10,000 | AST, CH37, (HKT) | 0.6677 | 80.33% |

Table 2: Results for AFP (kNN; $\alpha = 0.2$)

| k | Evaluations | Selected features | Best fitness | Avg. test accuracy |
|---|---|---|---|---|
| 3 | 1,000 | (AGE), AST, CH37, GT37, HB, (HKT), (PLT), (TBIL), (WBC) | 0.8574 | 78.38% |
| | 10,000 | AST, CH37, (HKT), (WBC) | 0.8226 | 77.93% |
| 5 | 1,000 | AST, (BUN), CH37, GT37, (HB), (TBIL) | 0.7822 | 79.02% |
| | 10,000 | AST, (BUN), CH37, GT37, (HB) | 0.7687 | 79.30% |
| 10 | 1,000 | AST, (BUN), CH37, (HB) | 0.7149 | 80.87% |
| | 10,000 | AST, (BUN), CH37, (HB) | 0.7149 | 80.03% |

Table 3: Results for AFP (ANN; $\alpha = 0.2$)

| Evaluations | Selected features | Best fitness | Avg. test accuracy |
|---|---|---|---|
| 1,000 | AST, CH37, (HKT) | 0.6482 | 82.54% |
| 10,000 | AST, CH37 | 0.6475 | 82.49% |

Table 4: Results for AFP (linReg; $\alpha = 0.5$)

| Evaluations | Selected features | Best fitness | Avg. test accuracy |
|---|---|---|---|
| 1,000 | AST, (CH37), (HB), (PLT) | 0.4636 | 79.24% |
| 10,000 | AST | 0.4556 | 78.91% |

Table 5: Results for AFP (kNN; $\alpha = 0.5$)

| k | Evalu-ations | Selected features | Best fitness | Avg. test accuracy |
|---|---|---|---|---|
| 3 | 1,000 | AST | 0.5872 | 75.90% |
| | 10,000 | AST | 0.5872 | 75.90% |
| 5 | 1,000 | AST | 0.5289 | 77.86% |
| | 10,000 | AST | 0.5289 | 77.86% |
| 10 | 1,000 | AST, (CH37), (HKT), (WBC) | 0.5385 | 79.93% |
| | 10,000 | AST | 0.5055 | 78.84% |

Table 6: Results for AFP (ANN; $\alpha = 0.5$)

| Evalu-ations | Selected features | Best fitness | Avg. test accuracy |
|---|---|---|---|
| 1,000 | AST | 0.4384 | 79.53% |
| 10,000 | AST | 0.4384 | 79.53% |

### 5.3. Test Results for CA 15-3

Table 7: Results for CA 15-3 (linReg; $\alpha = 0.2$)

| Evalu-ations | Selected features | Best fitness | Avg. test accuracy |
|---|---|---|---|
| 1,000 | (AGE), AST, (CH37), (HKT), (KREA), (PLT), | 0.6867 | 80.73% |
| 10,000 | AST, CH37 | 0.6677 | 80.33% |

Table 8: Results for CA 15-3 (kNN; $\alpha = 0.2$)

| k | Evalu-ations | Selected features | Best fitness | Avg. test accuracy |
|---|---|---|---|---|
| 3 | 1,000 | (AST), (BUN), (CEOA), (CMOA), (CNEA), (HKT), LD37 | 0.8979 | 67.90% |
| | 10,000 | (AST), (BUN), (CEOA), (CNEA), (HKT), LD37 | 0.8847 | 67.73% |
| 5 | 1,000 | (CEOA), (CMOA), (CRP), (GT37), (HKT), LD37, (WBC) | 0.8149 | 70.44% |
| | 10,000 | (CEOA), (CMOA), (CRP), (HB), LD37 | 0.8104 | 70.23% |
| 10 | 1,000 | (BUN), (CEOA), (CMOA), (CRP), (HB), (HKT), LD37 | 0.7568 | 72.01% |
| | 10,000 | (BUN), (CEOA), (CMOA), (CRP), (HB), (HKT), LD37 | 0.7522 | 71.37% |

Table 9: Results for CA 15-3 (ANN; $\alpha = 0.2$)

| Evalu-ations | Selected features | Best fitness | Avg. test accuracy |
|---|---|---|---|
| 1,000 | (AST), (CBAA), (CNEA), HB, LD37, (GT37), (PLT), (SEX) | 0.7030 | 73.91% |
| 10,000 | (AST), (CNEA), (GT37), (HB), LD37 | 0.6793 | 73.18% |

Table 10: Results for CA 15-3 (linReg; $\alpha = 0.5$)

| Evalu-ations | Selected features | Best fitness | Avg. test accuracy |
|---|---|---|---|
| 1,000 | AST, (CH37), (HB), (PLT) | 0.4636 | 79.24% |
| 10,000 | AST | 0.4556 | 78.91% |

Table 11: Results for CA 15-3 (kNN; $\alpha = 0.5$)

| k | Evalu-ations | Selected features | Best fitness | Avg. test accuracy |
|---|---|---|---|---|
| 3 | 1,000 | (CMOA), (HKT), LD37 | 0.6196 | 65.80% |
| | 10,000 | (CMOA), (HKT), LD37 | 0.5878 | 65.45% |
| 5 | 1,000 | (CMOA), (HKT), (CRP), (LD37) | 0.5860 | 65.97% |
| | 10,000 | (CMOA), (HKT), (CRP), (LD37) | 0.5731 | 65.23% |
| 10 | 1,000 | (CEOA), (CEA), (GT37), LD37, (RBC) | 0.5602 | 70.46% |
| | 10,000 | (CEOA), (GT37), LD37, (RBC) | 0.5488 | 70.21% |

Table 12: Results for CA 15-3 (ANN; $\alpha = 0.5$)

| Evalu-ations | Selected features | Best fitness | Avg. test accuracy |
|---|---|---|---|
| 1,000 | (LD37), (AST), (HB) | 0.4863 | 70.94% |
| 10,000 | AST, HB | 0.4674 | 72.35% |

## 6. CONCLUSIONS AND OUTLOOK

From the tables given in Section 5 we see that in most cases rather small feature sets are sufficient for achieving satisfying classification results: The test classification accuracies documented here are in most cases comparable to those summarized in (Winkler, Affenzeller, Jacak, and Stekel 2010); as expected, the best classifiers are here produced using ANNs. Obviously, setting $\alpha$=0.5 leads to very strong parsimony pressure, whereas setting $\alpha$=0.2 seems to be the better choice as it leads to better classification results using only comparably small feature sets. Furthermore, we also see that the test results documented after 1,000 evaluations are in many cases better than those documented after 10,000 evaluations.

On the one hand, future work should deal with the identification of feature sets and models for other tumor markers which are already available in the data based used in this research work; on the other hand, the authors will also try to identify estimation models for tumor diagnoses based on tumor marker information as well as standard blood parameters.


## 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.

Duffy, M. J. and Crown, J., 2008. A personalized approach to cancer treatment: how biomarkers can help. *Clinical Chemistry*, 54(11), pp. 1770–1779

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.

Niv, Y., 2008. Muc1 and colorectal cancer pathophysiology considerations. *World Journal of Gastroenterology*, 14(14), pp. 2139–2141.

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., 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 Upper Austria University of Applied Sciences (UAS), Campus Hagenberg.

**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** is a research associate at the UAS Research Center Hagenberg. His research interests include genetic programming, machine learning, and data mining and knowledge discovery. Currently he works on practical applications of data-based modeling methods for complex systems within *Heureka!*.

**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.