YamiPred: A novel evolutionary method for predicting pre-miRNAs and selecting relevant features

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2014.2388227, IEEE/ACM Transactions on Computational Biology and Bioinformatics

YamiPred: A novel evolutionary method for predicting pre-miRNAs and selecting relevant features

Dimitrios Kleftogiannis, Konstantinos Theofilatos, Spiros Likothanassis, and Seferina Mavroudi

Abstract—MicroRNAs (miRNAs) are small non-coding RNAs, which play a significant role in gene regulation. Predicting miRNA genes is a challenging bioinformatics problem and existing experimental and computational methods fail to deal with it effectively. We developed YamiPred, an embedded classification method that combines the efficiency and robustness of Support Vector Machines (SVM) with Genetic Algorithms (GA) for feature selection and parameters optimization. YamiPred was tested in a new and realistic human dataset and was compared with state-of-the-art computational intelligence approaches and the prevalent SVM-based tools for miRNA prediction. Experimental results indicate that YamiPred outperforms existing approaches in terms of accuracy and of geometric mean of sensitivity and specificity. The embedded feature selection component selects a compact feature subset that contributes to the performance optimization. Further experimentation with this minimal feature subset has achieved very high classification performance and revealed the minimum number of samples required for developing a robust predictor. YamiPred also confirmed the important role of commonly used features such as entropy and enthalpy, and uncovered the significance of newly introduced features, such as %A-U aggregate nucleotide frequency and positional entropy. The best model trained on human data has successfully predicted pre-miRNAs to other organisms including the category of viruses.

Index Terms—Classifier design and evaluation, Evolutionary computing and genetic algorithms, Feature evaluation and selection, SVM, GA, pre-miRNA prediction

1 Introduction

The discovery of miRNA genes was a breakthrough for conventional molecular biology and changed drastically the way we study and understand the underlying cellular processes [1]. Typically, miRNA genes are small in length (approximately 22 nucleotide long) and stable non-coding molecules, which play a significant role in gene regulation. By 2014, there have been identified more than 1880 human miRNAs, approximately 3500 in other mammals and many in flies, plants and viruses [2]. The active miRNA molecules (mature miRNAs) are produced by a number of biochemical reactions catalyzed by enzymes Drosha and Dicer via a procedure, which is called miRNA biogenesis [3], [4]. The biogenesis of miRNA starts in nucleus with the primary-RNA (pri-RNA) transcript that is further cleaved by Drosha to a shorter molecule, which is called precursor miRNA (pre-miRNA). The pre-miRNA is then transported to cytoplasm and further cleaved by Dicer to a double stranded hairpin-like molecule. One strand forms the mature miRNA, which is the active molecule that takes part in gene regulation processes by targeting messenger RNA (mRNA) transcripts. Typically, miRNAs interact with the RNAi Induced Silencing Complex (RISC), which has the ability to recognize and repress mRNA target genes [5].

The mRNA binding positions (called seeds) usually are located to the 3′ un-translated regions (UTR) of mRNAs but recent studies indicate that 5′ UTR binding, gene coding sequence (CDS) binding or binding to promoters is also possible [6]. It is estimated that at least 30% of all transcripts are regulated by miRNAs [7] and miRNAs layer of gene regulation is present in many cellular processes such as development, proliferation and apoptosis [8]. Gene expression studies indicate that abnormal miRNA homeostasis is linked to many diseases. Supplementary Table 1 illustrates important diseases related to abnormal miRNA machinery. Moreover, recent advances in biomedical research established new knowledge for the regulatory mechanisms of genes and genomes [9]. Consequently, the effective identification of miRNA genes remains an important bioinformatics problem and a crucial

- *Dimitrios Kleftogiannis is with the King Abdullah University of Science and Technology (KAUST), Computer Science and Mathematical Sciences and Engineering Division, Thuwal, Saudi Arabia. E-mail: dimitrios.kleftogiannis@kaust.edu.sa*
- *Konstantinos Theofilatos is with the Department of Computer Engineering and Informatics, University of Patras, Greece. E-mail:theofilk@ceid.upatras.gr*
- *Spiros Likothanassis is with the Department of Computer Engineering and Informatics, University of Patras, Greece. E-mail: likothan@ceid.upatras.gr*
- *Seferina Mavroudi is with the Department of Computer Engineering and Informatics, University of Patras, Greece. E-mail: mavroudi@ceid.upatras.gr*

step for developing more sophisticated therapeutic strategies [10]. The very first miRNAs and their targets were discovered experimentally through classical genetic techniques. However, experimental identification of miRNA genes has many drawbacks such as expensive laboratory reagents, time consuming experiments and low specificity. To overcome these hurdles computational techniques have been proposed. Computational approaches are classified into comparative and non-comparative [11], [12]. The former usually applies filtering criteria to identify miRNA gene candidates based on conservation among close species. The most representative comparative methods are MiRScan [13], MirAlign [14], MirCheck [15]. The latter category of non-comparative techniques is based on Computational Intelligence (CI) algorithms. Classification techniques such as Naïve Bayes Classifiers (NBC), Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Random Forests (RF) have successfully been applied. Among them, TripletSVM [16], miPred [17] and microPred [18] achieve very high performance by applying the SVM classifier. Lately a method that uses the SVM classifier combined with a simple Genetic Algorithm (GA) that identifies optimal feature subsets under the wrapper setting was proposed [19]. Despite the promising results in terms of performance, in-depth study of all the aforementioned techniques revealed some limitations. Specifically, the class imbalance problem, fine-tuning of classification parameters, over-fitting issues, tradeoff between classification performance and interpretability of results and selection of meaningful features are open problems for further consideration and optimization.

In this work, we propose YamiPred (**Y**et **a**nother **mi**RNA **pred**ictor), an embedded classification method that combines the efficiency and robustness of SVMs with GAs for feature selection and parameters optimization. YamiPred was trained using human pre-miRNA sequences. The feature vector includes state-of-the-art thermodynamical, structural and sequence features, plus 10 newly introduced characteristics, which have been proposed in the literature. Experimental results show that YamiPred outperforms state-of-the-art CI approaches and the prevalent SVM-based predictors in terms of classification performance and simplicity of the extracted classifiers. These advantages are attributed to the elegant way of dealing with the class imbalance problem, slow convergence and interpretability through a simple mechanism for selecting the ratio of positive and negative samples. A further contribution of YamiPred is a newly introduced problem-specific fitness function, which achieves more balanced classification performance between majority and minority class and simultaneously forces the algorithm to search for simpler classification models in terms of input features and model characteristics (i.e., support vectors).

Furthermore, the embedded feature selection component revealed characteristics such as minimum free energy, entropy and enthalpy that are fingerprints for different categories of non-coding RNAs (ncRNAs are categories of RNA that are not translated to proteins). Surprisingly, features not used by previous models such as %A-U aggregate nucleotide frequency, positional entropy and sequence length found to be important for predicting pre-miRNAs.

The best prediction model obtained from human data predicted with satisfactory results miRNAs in ten other organisms. These results indicate that YamiPred achieves good generalization capabilities and captures relevant miRNA properties across different species. Finally, experimentation with the most frequently selected features reported much higher classification performance and identified the minimum number of real pre-miRNAs required for developing a robust predictor. Thus, YamiPred is applicable to genomes where very few real pre-miRNAs are known. All the above-mentioned contributions and experimental results convincingly demonstrate that YamiPred is substantially different compared to existing SVM-based miRNA predictors [16], [17], [18], [19] and that it can provide a useful complement to these existing models to aid performing the challenging task of predicting miRNAs.

2 Materials and Methods

2.1 Datasets

In order to distinguish between real pre-miRNAs and other pseudo hairpins, YamiPred was trained and evaluated using datasets comprised of both positive and negative samples. For comparison reasons, the data construction step is analogous to the one described by miPred [17] and it is adopted by many other studies [18], [20]. All pre-miRNA sequences were extracted from publicly available databases. During the learning phase YamiPred was trained using human pre-miRNAs. 1,600 miRNA precursors of *Homo sapiens* published in miRBase (August 2013) [2] were extracted with average length 84 nucleotides (nt), minimum 43 nt and maximum 154 nt. Similarly to miPred and microPred we did not perform filtering steps to exclude sequences with diverse folding structures or multiple loops.

We generated negative dataset following the methodology proposed by Triplet-SVM [16] that uses known protein-coding regions for generating negative samples. These protein-coding regions have verified functionality and consequently, it is a straightforward approach for labeling negative data. Protein-coding regions were downloaded from RefSeq genes (registry August 2013) [21] and by applying a non-overlapping sliding window we generated pseudo hairpins using the following filtering criteria: i) minimum free energy less than -15 kcal/mol; ii) lowest number of base pairs equals to 18; iii) no multiple loops. These filtering criteria mimic real pre-miRNA properties. We also added the state-of-the-art pool of pseudo-hairpins that contains 8,494 sequences with average length 85 nt, minimum 63 nt and maximum 120 nt. Moreover, the negative dataset was enriched with 754 sequences coming from known ncRNAs originally published in [22]. This dataset consists of annotated ncRNA molecules such as tRNAs, snoRNAs and snRNAs with average length 84 nt, minimum length 48 nt and maximum 548 nt. These additional ncRNA sequences enriched the pool of negative samples and enhanced the ability to discriminate real pre-miRNAs from other categories of ncRNAs. The final negative set contains 21,248 pseudo-miRNAs. During the learning phase, we chose the simple holdout approach to generate training and testing sets. Specifically, positive and negative samples were pooled together and we generated randomly two completely independent datasets of equal size, one for training and one for testing. The large number of samples ensures that holdout is as effective as cross-validation techniques and addresses effectively potentials to over-fitting [23].

Then, to study YamiPred's generality we conducted cross-species experiments with independent test sets from several organisms. We chose species relatively distant to human such as Aves and Rodentia and the most representative mammalians starting from Carnivora and Laurasiatheria to other Primates and species very close to human such as gorillas and chimpanzees. Finally, to assess YamiPred's performance to a special category of miRNAs we downloaded all known viral miRNAs from miRBase repository. Species-specific pseudo-hairpin datasets were also generated for all the studied organisms using the aforementioned methodology [16]. We extracted 341 pre-miRNAs of *Equus caballus* and generated 517 pseudo hairpins, 323 pre-miRNAs of *Canis familiaris* and 520 pseudo hairpins, 500 pre-miRNAs of *Gallus gallus* and 2,000 pseudo hairpins, 85 of *Gorilla gorilla* miRNAs and 1,000 pseudo hairpins, 720 pre-miRNAs of *Mus musculus* and 3,000 pseudo hairpins, 88 precursors of *Pan paniscus and* 166 pseudo hairpins, 581 pre-miRNAs of *Pongo pygmaeus* and 1,356 pseudo hairpins, 662 *Bos Taurus* pre-miRNAs and 3,000 pseudo hairpins, 600 pre-miRNAs *of Pan troglodytes* and 176 pseudo hairpins and 237 known viral pre-miRNAs and 107 pseudo hairpins. All the datasets used for training and performance evaluation are publicly available along with the trained models for reproducing the results at the following repository (http://prlab.ceid.upatras.gr/microRNAdatasets/miRNA_datasets.rar).

2.2 Feature Set Description

Selecting an informative feature set is very important for the pre-miRNA prediction problem, as limited information is known about features that are able to distinguish between real miRNA and pseudo hairpins. Up to now various feature sets have been proposed, containing information about sequence, topology and structure. The earliest CI approaches such as Triplet-SVM proposed features computed from the sequence itself without including additional thermodynamical characteristics. miPred was the first method that proposes a representative feature set consisting of 29 attributes from various categories. MicroPred and its refined version [24] extended miPred's feature set to 45 attributes. The features proposed by miPred and microPred have shown great discriminative power and they have been adopted by many other methods [25], [26]. Here, we included the above-mentioned features and we added some new characteristics that characterize efficiently the broad class of ncRNAs such as snRNAs and rRNAs. We did not incorporate the left-triplet coding scheme proposed by Triplet-SVM because we computed single (mono), di and aggregate nucleotide frequencies that capture significant information deriving from the sequence itself. Furthermore, features that require phylogenetic filtering, alignment among species and expression profiles characteristics were not included due to the lack of sufficient data samples and the problem of missing values. The final feature set consists of 58 features. We adopted the same symbols used in miPred and microPred and the feature vector was computed using software written by the authors. Table 1 demonstrates the attributes categories. A more detailed description about the features can be found in Supplementary Materials.

2.3 YamiPred's Framework

2.3.1 Method Overview

YamiPred is an embedded classification framework, which combines an adaptive GA with an SVM classifier. Figure 1 provides a schematic representation of the method. The SVM algorithm is the most popular kernel method, due to its theoretical underpinnings and strong empirical performance on a wide variety of classification tasks [27]. It is a state-of-the-art classification technique that provides accurate models because it captures complex non-linearities in data. Furthermore, its strong mathematical background reassures high generalization performance [28]. When using SVMs with Radial Basis Function (RBF), it is necessary to select the best feature subset for the classifier and the optimal set of parameters (regularization parameter *C* and the RBF's bandwidth *gamma*). In order to optimize both, we use an embedded method for feature selection and parameter optimization based on an Adaptive Genetic Algorithm (AGA). Genetic Algorithms are heuristic optimization algorithms inspired by the principle of natural selection [29]. GAs deal with large and complicated search spaces (since they are guided by a problem-specific fitness function) and maintain a more global search strategy meaning that they are less likely to get trapped in local optima solutions compared with other search algorithms. GAs perform informed search that exploits and explores simultaneously the search space. The search strategy of a GA starts with a population of candidate solutions, called *chromosomes,* which is evolved and optimized via a number of evolutionary cycles and genetic operations. *Chromosomes* consist of *genes*, which are the parameters for optimization. For every iteration (called *generation*), a problem-specific fitness function is used to evaluate each chromosome, measuring the quality of the corresponding solution, and those that achieve the highest score (i.e., fitness value) are selected to survive to the next generation. This evolutionary process is continued until some user-defined termination criteria are met.

2.3.2 Crossover and Mutation operators

In YamiPred's implementation *chromosome* comprises of *genes* that encode the best feature subset and *parameter genes* that encode the best choice of parameters. Since YamiPred's optimization procedure is governed by the feature selection problem, which is binary (i.e., feature is present or not), for simplicity we used binary encoding for chromosome representation. The size of the initial population was set to 20 chromosomes. For the crossover operator, we used two-point crossover with rate of 0.9, which is considered, typical crossover rate for many GA applications. Note that, higher crossover rates are preferable in GAs [30].

The mutation operator favours exploitation over exploration. In the first generations of the algorithm it is preferable to explore a wider search space (exploration) while in the last generations it is preferable to search locally the most promising areas of the search space (exploitation). Many solutions have been proposed in the literature for adaptive mutation operators ranging from self-adaptive ones [31], deterministic approaches and gene based adaptive mutation operators [32]. Deterministic approaches as proposed in [33] are the most suitable solutions for difficult combinatorial problems where the search space is large and complex. In YamiPred we used an adaptive mutation rate approach that extends the ideas presented in [33] to simulate the above-mentioned exploration-exploitation property. YamiPred's mutation rate starts with a high mutation probability that gradually decreases to switch from global to local search. The mutation rate is computed using equation (1):

$$Pm(n) = 0.2 - n \cdot \frac{0.2 - \frac{1}{P_S}}{\mathrm{MAX_G}} \quad (1)$$

where n is the current generation, P_S is the size of the population and MAX_G is the maximum generation specified by the termination criteria. In order to avoid getting trapped into local optima the deterministic adaptive mutation rate was extended to introduce instant increments in the mutation rate when the possibility of stagnation was increased. The mean similarity of every individual with the best individual of the population was measured at every generation. If the mean was larger than 90% then the mutation probability was increased by a factor given by relation (2) instead of being decreased.

$$\frac{0.2 - \frac{1}{p_s}}{MAX_G} (2)$$

2.3.3 Selection scheme and Fitness Function

In YamiPred's selection scheme we applied rank based roulette wheel selection. In this scheme a fitness value, equal to the rank in the population, and not equal to the actual objective value, is assigned for each individual and thus the highest ranked individual has the highest probability to be selected in the next generation. This probability was calculated based on the following equation:

$$Pi = \frac{2 * Rank}{N * (N + 1)} \quad (3)$$

where N is the number of individuals in the population. Selection based on objective value can promote premature convergence when there is a large difference between these values, and this is the main reason for utilizing Pi values instead of the actual objective values. Following this selection scheme, the proposed evolutionary algorithm forces the population to areas of better solutions while reducing the possibilities of getting trapped into local optima [29].

YamiPred's selection scheme also utilized elitism that forces the best solutions of every population to be selected at least one time in the next population. Extensive usage of elitism sometimes leads to premature convergence. However, YamiPred's implementation that keeps only the optimal solution in the next geneneration and utilises Pi values instead of the actual objective values aleviates the problem. In the examined optimization problem we have defined two important sub-objectives. The first and most important sub-objective is to maximize classification performance.

The second sub-objective is to develop a relatively simple classification model that uses compact feature subsets and few support vectors. Since our problem's sub-objectives are grouped into two main contradictory goals, a single objective optimization method was used defining a problem-specific fitness function as in (4).

$$Fitness = a \cdot Accuracy + b \cdot GeometricMean - c \cdot Features - d \cdot SupportVectors \quad (4)$$

where *Accuracy* is SVM's accuracy, GeometricMean is the geometric mean of sensitivity and specificity, *features* is the number of selected features and SupportVectors is the number of support vectors included in the trained SVM model. The proposed fitness function balances classification performance, complexity of the feature set and classification model's complexity. We chose Accuracy because it is a general performance metric that measures proximity of the observed values to the true values and it is efficient with balanced data. We also chose Geometric Mean that captures the combined effect of the other two important performance metrics and it remains efficient with imbalanced datasets [34]. Furthermore, integrating the number of support vectors into the optimization process demonstrates novelty and differentiates YamiPred from other wrapper-based tools like the one presented in [19]. Overall, terms *Features* and *SupportVectors* have negative effects in the fitness value because YamiPred's goal is to produce a simple classifier (without losing classification performance) that has higher generalization capabilities and produces interpretable results.

Parameters a, b, c and d in equation (4) are user specified weights, which incorporate prior knowledge about the significance of a model's sub-objectives. Specifically, classification Accuracy and Geometric mean are the most significant sub-objectives whereas the number of selected

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2014.2388227, IEEE/ACM Transactions on Computational Biology and Bioinformatics

features is less significant and the number of support vectors appears as the least significant sub-objective. Based on this ordering we selected the sum of weights for sub-objectives related to classification performance to be two times bigger than the sum of weights for sub-objectives related to model's simplicity. Finally, we assign the following values for the constants parameters: a=0.5, b=0.5, c=0.001 and d=0.0001. Note that we did not conduct experiments to optimize these values as this may lead to over-fitting.

2.3.4 Termination Criteria

YamiPred's termination criterion is the maximum number of 150 generations to be reached combined with a termination "flag" that stops the process when the population is deemed as converged. The population is deemed as converged when the average fitness across the current population is less than 5% away from the best fitness of the current population. During the experimentation process we observed that the convergence criterion was satisfied between the 100th and the 150th generation.

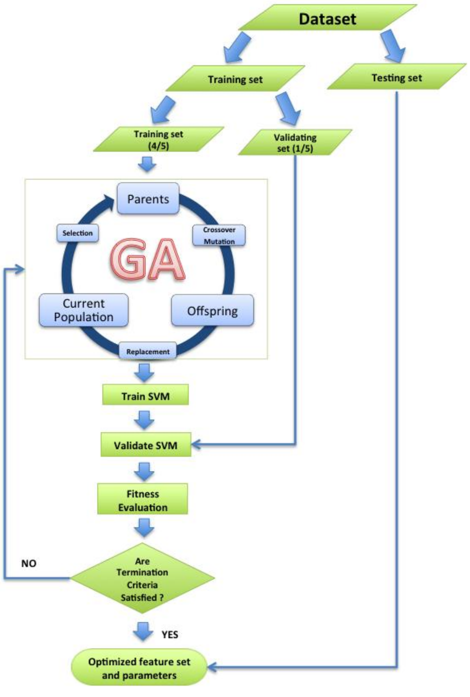

Fig. 1. YamiPred's workfow

2.3.5 Software availability and time efficiency

YamiPred's implementation held in Matlab R2009b and source codes are available at (http://prlab.ceid.upatras.gr/microRNAdatasets/codes.rar) Concerning the time efficiency of the training phase, it is quite time consuming as it performs multiple SVM trainings and searches thoroughly the search space (requiring on average 16 hours and 15 minutes for every experiment using a conventional laptop equipped with an Intel Core i5 processor at 1.7 GHz). However, this is an offline procedure. More important in terms of time efficiency is the minimization of the time required to apply the trained model to new candidate sequences. To illustrate the time efficiency of YamiPred we measured the average classification time for processing 100 human sequences. For this task, YamiPred needs 0.0013 seconds for prediction.

Table 1

Feature Set Description

Category	Description	Number
Dinucleotide Frequencies	%XY such that X,Y e Σ[A,C,G,U]	16
Aggregate Dinucleotide Frequency	%G+C ratio	1
Folding Measures	Various topological and sequential identifiers normalized by length	4
Minimum Free Energy indices	Adjusted Minimum Free Energy normalized by various identifiers of the secondary structure	4
Topological Descriptor	RAG	1
RNA fold features	Vienna RNA package [40]	4
Una Fold features	Una Fold package [41]	6
Base pairs related features	Number of significant base pairs normalized by length and total number of stem loops	8
Statistical Features	The statistical Z-score of the folding measures	4
Newly introduced attributes	Additional features from various categories applicable to other ncRNA molecules	10

3 Results and Discussion

3.1 Dealing with the class imbalance problem

Pre-miRNA classification problems are imbalanced, meaning that there are far fewer data from the class of interest (pre-miRNAs) than from the negative class (pseudo-hairpins) [34]. The explanation is intuitively simple because in cells the quantity of molecules, which are not pre-miRNAs, and fold into a miRNA-like shape is larger than the real miRNA genes. The earliest approaches such as Triplet-SVM [16] and miPred [17] reported the problem and balanced datasets manually with respect to minority class. More recent methodologies applied a variety of different techniques for dealing with the problem. Striking examples are PlantMiRNAPred [26], which applied sampling according to the sample distribution in the positive/negative groups and microPred [18], which tested a variety of class-imbalance learning techniques including random over/under sampling [35] and SMOTE [36]. Note that many state-of-the-art methods do not select the ratio between positive and negative samples so as to maximize certain performance criteria and to this extent they do not get feedback from classifiers. In the present work, we used an internal approach, which deals with the problem as part of the learning phase of YamiPred. To do so, the learning phase was repeated multiple times with various positive to negative sample rates [37]. Table 2 presents the results of the proposed method using 5-fold cross validation applied on the testing set. These experimental results suggest the usage of four times more negative samples than positive achieves the highest fitness value (using equation 4). Note that, this finding was not induced using the performance of the proposed model in the testing set, as this procedure would have lead to overfitting. The testing set was used only for the final evaluation of the proposed prediction model.

3.2 Comparison with existing methods

To estimate the effectiveness of YamiPred and quantify the contribution of GA in feature selection and the parameter optimization process we performed experiments using YamiPred and two YamiPred variants. The first one (denoted as SVM_GA_v1) utilized the GA to optimize only feature subsets using default SVM parameters C and gamma (1 and 1, respectively). The second variant (denoted as SVM_GA_v2) used all the available features and applied the GA optimizer for optimizing SVM parameters only. To provide better cross benchmarking results YamiPred and its variants were further compared against classical CI approaches such as NBC, K- Nearest Neighbors (KNN) and RF. These CI algorithms along with SVM are the most commonly used classification techniques in bioinformatics.

Table 2

YamiPred's performance using various positive to negative samples rate

Positive to Negative Samples Rate	Accuracy	Geometric Mean	Selected Inputs	Support Vectors	Fitness Value
(1:1)	0.909 ±0.004	0.909 ±0.003	25±3	**828.41** **±122.73**	1.7205 ±0.002
(1:2)	0.923 ±0.005	0.918 ±0.005	24±2	860.64 ±135.18	1.738 ±0.001
(1:3)	0.932 ±0.004	0.916 ±0.005	23±2	870.88 ±153.27	1.739 ±0.014
(1:4)	0.931 ±0.005	**0.924** **±0.004**	20±2	879.25 ±101.39	**1.740** **±0.006**
(1:5)	0.929 ±0.004	0.915 ±0.007	**18±1**	913.55 ±170.71	1.736 ±0.002
(1:6)	**0,936** **±0,003**	0,900 ±0,004	23±2	943,42 ±201,52	1,719 ±0,001

Tuning of their internal parameters was performed using trial and error experiments on the training set and using geometric mean as the performance indicator. Due to the stochastic nature of these approaches all experiments were executed twenty times and their mean performances in the testing set are demonstrated in Table 3. Comparing YamiPred with its variants (SVM_GA_v1 and SVM_GA_v2) we observe that SVM combined with GA for parameter optimization and feature selection achieves the best results and it is the most effective. Thus, the contribution of the GA search strategy in the embedded setting is valuable. We also observe that SVM-GA is superior to the prevalent CI classification models. Since some differences in the performance are small, we applied a test that quantifies practically these differences [38]. Considering all the performance metrics in the comparison, YamiPred always appears as the best method. The detailed results are presented in Supplementary Materials. These differences were also found to be statistically important when applying statistical t-test for independent samples with 95% level of significance (normality of the data was ensured using Shapiro Wilks test). We also observed that RF achieves slightly better accuracy than SVM-GA and much higher sensitivity. However, this improvement in the sensitivity metric was achieved with an important degradation in specificity leading to an imbalanced classifier, which presented lower geometric mean than the proposed method.

Except for the prevalent CI algorithms, YamiPred was also compared with the state-of-the-art SVM methods miPred and microPred. Table 4 presents the results. YamiPred again is the best performing method. However, since, the differences in performance are small, we applied the previous test [38] to get practical insights about their relative performance. We found again that YamiPred outperforms both competitors. In addition, YamiPred's fitness value was proved to be significantly superior to the other fitness values according to t-test (normality of the data was checked using Shapiro Wilks test).

3.3 Selecting relevant features

Performing closer examination of the developed classification models we discovered that YamiPred resulted in different feature subsets for every execution. This result was expected since some of the features, which constitute the initial feature set, share mutual information. This implies that there exists more than one combination of features that maximize classification performance. In addition, a subset of features was in most cases consistent and fulfilled an important property, which is called stability [39]. The features which were selected by YamiPred in more than 80% of execution runs were the following 8 attributes: Dinucleotide Frequencies AG and AU, Normalized base pair distance (dD), Positional Entropy (PosEntropy), Normalized Ensemble free energy (EAFE), Enthalpy normalized by the length of the sequence (dH/L), Melting temperature normalized by the length (Tm/L) and the length of the sequence itself. When these features were used as inputs to a SVM model they achieved 90,32% accuracy and 88,54% geometric mean. Despite the fact that this classification performance was not as high as the performance reported in Table 4, it is interesting to study these features and understand the mechanisms that govern whether a hairpin is a pre-miRNA or not. Based on the reported 8 features we found that sequence information and other thermodynamical characteristics such as minimum free energy, entropy, enthalpy and melting temperature act like a fingerprint for the different categories of ncRNAs.

Table 3

Comparison with state-of-the-art CI techniques

Classification Method	Accuracy	Sensitivity	Specificity	Geometric Mean
NBC	0.914±0.003	0.943±0.003	0.796±0.012	0.867±0.006
KNN	.908±0.005	0.970±0.122	0.657±0.023	0.798±0.009
RF	**0.937±0.004**	0.979±0.002	0.765±0.002	0.865±0.008
YamiPred	0.932±0.005	0.937±0.008	**0.912±0.012**	**0.924±0.004**
SVM-GA_v1 (only feature selection)	0,931±0,003	**0,945±0,004**	0,875±0,002	0,909±0,001
SVM-GA_v2 (only parameter selection)	0,930±0,001	0,940±0,002	0,892±0,005	0,916±0,003

Table 4

Comparison with state-of-the-art SVM predictors

	YamiPred	miPred	microPred
Accuracy	**0.932±0.005**	0.927±0.007	0.926±0.003
Sensitivity	**0.937±0.008**	0.934±0.007	0.934± 0.001
Specificity	**0.912±0.012**	0,899±0.018	0.892±0.015
Geometric Mean	**0.924±0.04**	0.916±0.011	0.913±0.009
Selected Inputs	**20±2**	29±0	21±0
Support Vectors	**879.25±101.39**	1129.333±267.95	1115.33±55.01
Fitness Value	**1.74±0.001**	1.702±0.011	1.707±0.007

The best run of YamiPred in terms of classification performance identified a feature subset that contains 20 features which are the following: aggregate nucleotide frequency A+U, dinucleotide frequencies AG, AU, CU, GA, UU, Minimum Free Energy Index 4 (MFEI4), Positional Entropy (PosEntropy), Normalized Ensemble free energy (EAFE), Frequency of the MFE structure (Freq), Enthalpy normalized by the length of the sequence (dH/L), Melting temperature (Tm), Melting temperature normalized by length (Tm/L), Normalized base-pair count by length ,|G-C|/L, Normalized average base pairs by number of stem loops (A-U)/stems, (G-U)/stems, the length of the sequence (Len), Centroid energy normalized by length (CE/L), Statistical Z-scores zG and zSP.

With a more thorough analysis on this optimal set of features we found that combination of features derived from different categories lead to better separation between real pre-miRNAs and other molecules. The final selected subset contains features from all the available categories presented in Table 1. Sequence information is captured by various nucleotide frequencies. Thermodynamical characteristics computed by Vienna RNA package [40] and UnaFold [41] provide a solid representation of the RNA structure. Second, our findings are in agreement with previous reports that linked different characteristics with biochemical properties and the secondary structure of molecules [42], [43]. Compared to miPred's and microPred's most relevant feature subsets we agree on the selection of 8 features and we conclude that those introduced by microPred are more suitable for describing the hairpin stem-loop while miPred's features give a better representation of the secondary structure of miRNAs. Note that YamiPred's most frequent set and best performing set do not agree with findings reported by the method presented in [19]. In contrast to those methods our model selected the %A-U aggregate nucleotide frequency instead of %G+C and enthalpy instead of entropy. A possible explanation comes from recent evidence that link the A-U content with stability of mRNAs [44]. Also, as expected and illustrated in Supplementary Material there is a clear relation between enthalpy, entropy, minimum free energy and temperature. Regarding the newly introduced features, YamiPred chose 5 out of 10 attributes. Thus, our initial hypothesis that more general characteristics, applicable to the broad class of ncRNAs, have higher discriminatory power has been validated.

In summary, YamiPred achieves higher classification performance than the prevalent CI classifiers and SVM-based approaches while it finds a compact feature subset. Thus, it is capable of reducing the problem's dimensionality and producing interpretable results without sacrificing the performance.

3.4 Predicting miRNAs in other organisms

In this section we study the generalization capabilities of YamiPred and we compare it to miPred and microPred. For this purpose we applied the best-trained models (using human datasets) of the competitors to ten datasets coming from several organisms. Figure 2 presents the classification results. We found that YamiPred's best-trained model achieves very high performance in almost all of tested cases. Specifically, we found that YamiPred always achieves better generalization ability than miPred and in 8 out of 10 cases better generalization than microPred. The higher classification performance achieved by YamiPred provides strong evidence that pre-miRNAs among various eukaryotic species and viruses share similar sequence, thermodynamical and structural properties. However, in some cases, such as the *Bos Taurus* dataset, the classification performance was lower than in other organisms. To shed light on this artifact we trained a model with *Bos Taurus* data, we measured the classification performance and studied the differences between the selected features subset and those selected from human. The specie-specific trained model further improved the performance of predicting *Bos Taurus* miRNAs by achieving geometric mean equal to 89.75% instead of 83.91% we previously achieved. Regarding the selected features, YamiPred selected 11 (%A+U, AU, CU, UU, MFEI4, EAFE, dH/L, Tm/L, |G-C|/L, (G-C)/stems, ZG) out of the best 20 features and 4 (AU, EAFE, dH/L, Tm/L) out of 8 of the most stable features. Moreover, it included additional characteristics such as more dinucleotide frequencies (AC, GA, GG, CA, UG), statistical score ZP, topological factor dF and BP/AU, BP/GU which seem to differentiate *Bos Taurus* miRNAs properties. All the above illustrate a differentiation of the miRNA structural and sequential properties of this specific organism against all the others.

3.5 Further improving Yamipred's performance

Finally, in order to improve further Yamipred's classification performance we experimented with the most frequently selected features described in section 3.3. This subset consists of 8 attributes that were present in almost all runs of the GA feature selection process. Using this small feature subset, we transformed the original feature vector and we generated paired differences between individual features and all the other features coming from all the other data samples. In total, this transformation generates N*K features where N is the original number of features and K is the number of samples in the dataset. Then, in order to reduce dimensionality we applied a heuristic technique and we obtained the median of the differences plus two additional features generated by subtracting standard deviation to the median and by adding standard deviation to the median. Note, that this is a heuristic technique (similar to the heuristics that all feature selection techniques use) that provides a good approximation of the initial feature vector [45]. In total the transformed feature vector consists of 24 attributes coming from 8 original features. To test the classification performance the SVM classifier was trained with various ratios between positive and negative samples and we reported results using 2-fold cross validation. We used 50% of the data for training and the remaining 50% for testing. Note that for tuning the SVM parameters we applied a simple grid search technique on a portion of the training set equal to 30% [46]. Table 5 presents the classification results. It is apparent that extremely high classification performance was achieved by applying the proposed transformation to the most frequently selected features by the GA. Moreover, we did not observe any remarkable performance degradation using different ratios between positive and negative samples, meaning that the data transformation reduces the effects of the class imbalance problem. Surprisingly, in all of the cases perfect specificity was reported meaning that the YamiPred's improved variant is able to identify accurately negative examples. Then in order to assess the robustness of this technique we aimed at identifying the minimal number of positive samples that is sufficient for achieving very high classification performance. For this purpose we trained models initially with the total number of positive samples and progressively we removed positive samples from the training set. The performance was recorded using 5-fold cross validation and we present these results in Table 6. The overall performance decreases below 400 samples but in general remains high even when the number of positive samples is very small. In the extreme case using only 10 positive samples improved YamiPred reported Accuracy=99.65%, Sensitivity=100%, Specificity=75.4%, and Geometric Mean=86.49%. From all the above we conclude that YamiPred improved variant is a robust predictor that can be trained with success even when the number of known positive samples is very limited. This is attributed to the sophisticated feature selection process that revealed a minimal number of attributes and to the normalization technique used.

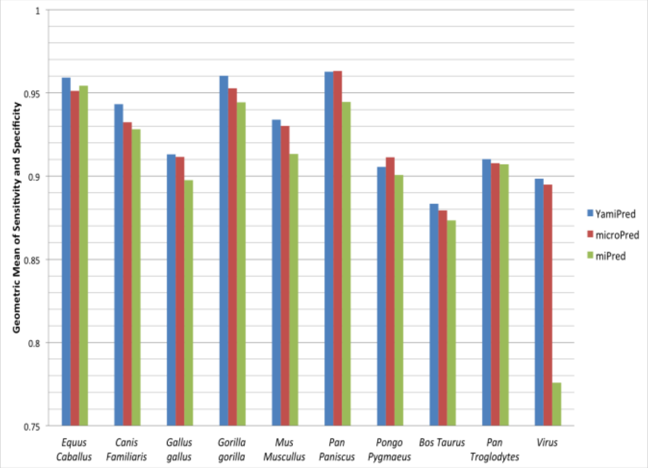

Fig. 1. Performance of different methods in various organisms

4 Conclusion

In this work we have presented YamiPred, another SVM-based miRNA predictor deploying the evolutionary characteristics of GAs. The GA optimizer maximizes the classification performance and integrates the feature selection phase as well as the parameter optimization step into the learning components of the methodology. The adoption of an objective function that includes number of support vectors and size of the selected features leads to a more general model that achieves very high performance in terms of accuracy sensitivity and specificity. Our experimental setup proved that we could handle effectively the obstacle of the class imbalance and moreover we can generalize the model to several organisms. Moreover, the properties of the selected features were studied and we reported a stable and relevant subset containing features with high discrimination power. Further experimentation with this subset of features has improved further the classification performance and revealed the minimum number of positive samples, which are required to implement a robust classifier. Indeed this fact makes YamiPred a robust miRNA predictor even for organisms for which a very small number of miRNA genes is known. Also, we have concluded that among different species the miRNA class shares common characteristics that act like fingerprints for this particular class of regulatory molecules. All the aforementioned properties fulfill the basic requirement for the development of effective and robust models and there is space for many future improvements. An interesting future area to explore is the incorporation of YamiPred to an integrated analysis framework that combines data from heterogeneous data sources into a cellular interaction network. In fact, recent projects based on ChIP-Seq data have performed a comprehensive analysis based on different functional arrays. For instance ENCODE project has established new knowledge regarding the distal regulatory element (enhancer and insulators) interactions and gene transcription. We plan to feed YamiPred with new problem-specific features, and apply it to other classes of regulatory RNAs including the class of enhancer RNAs (eRNA), which at the moment has unknown functionality. Furthermore, to deal with the class imbalance problem more effectively, our future plans involve the design of a new fitness function that uses a new evaluation metric called adjusted geometric mean (AGm) [47].

Table 5

Improved YamiPred performance using various positive to negative samples ratios

Positive to Negative	Accuracy	Sensitivity	Specificity	Geometric Mean
(1:1)	0.9951	0.9902	**1**	0.9951
(1:2)	0.9925	**0.9887**	**1**	0.9943
(1:3)	0.9983	0.9977	**1**	0.9988
(1:4)	0.9983	0.9978	**1**	0.9989
(1:5)	0.9979	0.9974	**1**	0.9987
(1:6)	**0.9986**	0.9984	**1**	**0.9992**

Table 6

Performance of improved YamiPred with progressively decreasing positive samples

Number of positive	Accuracy	Sensitivity	Specificity	Geometric Mean
1599	0.9951	0.9902	**1**	0.9951
1300	0.9956	0.9916	**1**	0.9958
1000	0.9942	0.9902	**1**	0.9950
700	0.9915	0.9873	**1**	0.9936
400	**0.9984**	0.9979	**1**	**0.9989**
100	0.9954	0.9972	0.9706	0.9836
50	0.9952	**1**	0.8627	0.9248
10	0.9965	**1**	0.7540	0.8649

5 References

[1] E. C. Lai, “microRNAs: runts of the genome assert themselves,” Curr. Biol., vol. 13, no. 23, pp. R925–936, Dec. 2003.

[2] S. Griffiths-Jones, “The microRNA Registry,” Nucleic Acids Res., vol. 32, no. Database issue, pp. D109–111, Jan. 2004.

[3] Y. Lee, K. Jeon, J.-T. Lee, S. Kim, and V. N. Kim, “MicroRNA maturation: stepwise processing and subcellular localization,” EMBO J., vol. 21, no. 17, pp. 4663–4670, Sep. 2002.

[4] G. M. Borchert, W. Lanier, and B. L. Davidson, “RNA polymerase III transcribes human microRNAs,” Nat. Struct. Mol. Biol., vol. 13, no. 12, pp. 1097–1101, Dec. 2006.

[5] S. M. Elbashir, W. Lendeckel, and T. Tuschl, “RNA interference is mediated by 21- and 22-nucleotide RNAs,” Genes Dev., vol. 15, no. 2, pp. 188–200, Jan. 2001.

6] Y. Tay, J. Zhang, A. M. Thomson, B. Lim, and I. Rigoutsos, “MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation,” Nature, vol. 455, no. 7216, pp. 1124–1128, Oct. 2008.

[7] A. Krek, D. Grün, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel, and N. Rajewsky, “Combinatorial microRNA target predictions,” Nature Genetics, vol. 37, no. 5, pp. 495–500, Apr. 2005.

[8] D. R. Hipfner, K. Weigmann, and S. M. Cohen, “The bantam gene regulates Drosophila growth,” Genetics,

vol. 161, no. 4, pp. 1527–1537, Aug. 2002.

9] M. Skipper, R. Dhand, and P. Campbell, "Presenting ENCODE," Nature, vol. 489, no. 7414, pp. 45–45, Sep. 2012.

[10] C. Li, Y. Feng, G. Coukos, and L. Zhang, "Therapeutic microRNA strategies in human cancer," AAPS J, vol. 11, no. 4, pp. 747–757, Dec. 2009.

11] N. D. Mendes, A. T. Freitas, and M.-F. Sagot, "Current tools for the identification of miRNA genes and their targets," Nucleic Acids Res., vol. 37, no. 8, pp. 2419–2433, May 2009.

[12] D. Kleftogiannis, A. Korfiati, K. Theofilatos, S. Likothanassis, A. Tsakalidis, and S. Mavroudi, "Where we stand, where we are moving: Surveying computational techniques for identifying miRNA genes and uncovering their regulatory role," J Biomed Inform, vol. 46, no. 3, pp. 563–573, Jun. 2013.

[13] L. P. Lim, N. C. Lau, E. G. Weinstein, A. Abdelhakim, S. Yekta, M. W. Rhoades, C. B. Burge, and D. P. Bartel, "The microRNAs of Caenorhabditis Elegans," Genes Dev., vol. 17, no. 8, pp. 991–1008, Apr. 2003.

[14] X. Wang, J. Zhang, F. Li, J. Gu, T. He, X. Zhang, and Y. Li, "MicroRNA identification based on sequence and structure alignment," Bioinformatics, vol. 21, no. 18, pp. 3610 –3614, Sep 2005.

[15] M. W. Jones-Rhoades and D. P. Bartel, "Computational identification of plant microRNAs and their targets, including a stress-induced miRNA," Mol. Cell, vol. 14, no. 6, pp. 787–799, Jun. 2004.

[16] C. Xue, F. Li, T. He, G.-P. Liu, Y. Li, and X. Zhang, "Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine," BMC Bioinformatics, vol. 6, p. 310, 2005.

[17] K. L. S. Ng and S. K. Mishra, "De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures," Bioinformatics, vol. 23, no. 11, pp. 1321–1330, Jun. 2007.

[18] R. Batuwita and V. Palade, "microPred: effective classification of pre-miRNAs for human miRNA gene prediction," Bioinformatics, vol. 25, no. 8, pp. 989–995, Apr. 2009.

[19] Y. Wang, X. Chen, W. Jiang, L. Li, W. Li, L. Yang, M. Liao, B. Lian, Y. Lv, S. Wang, S. Wang, and X. Li, "Predicting human microRNA precursors based on an optimized feature subset generated by GA-SVM," Genomics, vol. 98, no. 2, pp. 73–78, Aug. 2011.

[20] D. T.-H. Chang, C.-C. Wang, and J.-W. Chen, "Using a kernel density estimation based classifier to predict species-specific microRNA precursors," BMC Bioinformatics, vol. 9 Suppl 12, p. S2, 2008.

[21] F. Hsu, W. J. Kent, H. Clawson, R. M. Kuhn, M. Diekhans, and D. Haussler, "The UCSC Known Genes," Bioinformatics, vol. 22, no. 9, pp. 1036–1046, May 2006.

[22] E. S. Lander, L. M. Linton, B. Birren,..., and M. J. Morgan, "Initial sequencing and analysis of the human genome," Nature, vol. 409, no. 6822, pp. 860–921, Feb. 2001.

[23] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," BMC Bioinformatics, vol. 7, p. 91, 2006.

[24] R. Batuwita and V. Palade, "An improved non-comparative classification method for human microRNA gene prediction," in 8th IEEE International Conference on BioInformatics and BioEngineering, 2008. BIBE 2008, 2008, pp. 1–6.

[25] C.-H. Hsieh, D. T.-H. Chang, C.-H. Hsueh, C.-Y. Wu, and Y.-J. Oyang, "Predicting microRNA precursors with a generalized Gaussian components based density estimation algorithm," BMC Bioinformatics, vol. 11 Suppl 1, p. S52, 2010.

[26] P. Xuan, M. Guo, X. Liu, Y. Huang, W. Li, and Y. Huang, "PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNAs," Bioinformatics, vol. 27, no. 10, pp. 1368–1376, May 2011.

[27] D. P. Lewis, T. Jebara, and W. S. Noble, "Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure," Bioinformatics, vol. 22, no. 22, pp. 2753–2760, Nov. 2006.

[28] V.N. Vapnik, The nature of statistical learning theory. Springer, 2000

[29] Holland, Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence, Cambridge, Mass MIT Press, 1995.

[30] R. Colin and J. Rowe, Genetic algorithms: principles and perspectives: a guide to GA theory. Vol. 20. Springer, 2003

[31] T. Back , "Optimal Mutation Rates in Genetic Search", Proceedings of 5th International Conference on Genetic Algorithms, Morgan Kaufmann,1993

[32] T. Oong and N. Isa, Adaptive Evolutionary Artificial Neural Networks for Pattern Classification", IEEE Transcactions on Neural Networks, Vol 22, No. 11, November 2011

[33] D. Thierens , "Adaptive Mutation Control Schemes in Genetic Algorithms", Proceedings of Congress on Evolutionary Computing, IEEE 2002.

[34] R. Akbani, S. Kwek, and N. Japkowicz, "Applying Support Vector Machines to Imbalanced Datasets," in Machine Learning: ECML 2004, J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, Eds. Springer Berlin Heidelberg, 2004, pp. 39–50.

[35] G. M. Weiss, "Mining with Rarity: A Unifying Framework," SIGKDD Explor. Newsl., vol. 6, no. 1, pp. 7–19, Jun. 2004.

[36] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," J. Artif. Int. Res., vol. 16, no. 1, pp. 321–357, Jun. 2002.

[37] B. Raskutti and A. Kowalczyk, "Extreme Rebalancing for SVMs: A Case Study," SIGKDD Explor. Newsl., vol. 6, no. 1, pp. 60–69, Jun. 2004.

[38] A. Vargha and H.D. Delaney, "A Critique and Improvement of the "CL" Common Language Effect Size Statistics of McGraw and Wong", Journal of Educational and Behavioral Statistics, Vol.25, No.2 , p 101-132, 2000

[39] L. Yu, C. Ding, and S. Loscalzo, "Stable Feature Selection via Dense Feature Groups," in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, pp. 803–811, Aug 24-27 2008.

[40] I. L. Hofacker, "Vienna RNA secondary structure

server," Nucleic Acids Res., vol. 31, no. 13, pp. 3429–3431, Jul. 2003.

[41] N. R. Markham and M. Zuker, "UNAFold: software for nucleic acid folding and hybridization," Methods Mol. Biol., vol. 453, pp. 3–31, 2008.

[42] V. Moulton, M. Zuker, M. Steel, R. Pointon, and D. Penny, "Metrics on RNA secondary structures," J. Comput. Biol., vol. 7, no. 1–2, pp. 277–292, Apr. 2000.

[43] E. Freyhult, P. Gardner, and V. Moulton, "A comparison of RNA folding measures," BMC Bioinformatics, vol. 6, no. 1, p. 241, Oct. 2005.

[44] B. Shi, W. Gao, and J. Wang, "Sequence fingerprints of microRNA conservation," PLoS ONE, vol. 7, no. 10, p. e48256, 2012.

[45] K. Lim, Z. Li, K.P Choi and L. Wong "ESSNet: Finding consistent disease subnetworks in data with extremely small sample sizes ", Unpublished.

[46] B. Scholkopf, C. J. C. Burges and A. Smola, ¨ Advances in Kernel Methods: Support Vector Learning, Cambridge, MA: MIT Press, 1999.

[47] R. Batuwita and V. Palade, "Adjusted geometric-mean: a novel performance measure for imbalanced bioinformatics datasets learning," J Bioinform Comput Biol, vol. 10, no. 4, p. 1250003, Aug. 2012.

Dimitrios Kleftogiannis in 2009 received a Diploma in Computer Engineering from the Computer Engineering and Informatics Department University of Patras, Greece. In 2011 he received a Master degree in Informatics for Life Sciences from the Medical School of Patras, Greece. Currently he is a Phd candidate at King Abdullah University of Science and Technology (KAUST) under the supervision of professor Panos Kalnis. Dimitrios mostly works on knowledge discovery technologies with applications in bioinformatics.

Konstantinos Theofilatos in 2006 received his Diploma in Computer Engineering from Computer Engineering and Informatics Department University of Patras, Greece. From the same Department he received his master degree in 2009 and his Phd diploma in 2013. He is currently a post-doctoral researcher at the Department of Computer Engineering and Informatics in the University of Patras Greece.

Spiridon Likothanassis received the diploma degree in electrical engineering from the National Technical University (NTU), Athens, Greece, in 1980 and the Ph. D. degree in stochastic adaptive control from the Department of Computer Engineering and Informatics, University of Patras, Patras, Greece, in 1986. He is currently Professor and Director of the Pattern Recognition Laboratory, Department of Computer Engineering and Informatics, University of Patras.

Seferina Mavroudi graduated in 1998 from the Department of Electrical and Computer Engineering, School of Engineering of the Aristotles University of Thessaloniki. In 2000 she received a Master's degree from the European Postgraduate Program on Biomedical Engineering, organized by the Faculty of Medicine of the University of Patras, the Faculty of Mechanical Engineering and the Faculty of Electrical and Computer Engineering of the National Technical University of Athens, in collaboration with more than 20 European Universities. In the same program, in February of the year 2003 she completed her Ph.D. Seferina Mavroudi is a lecturer in the Department of Social Work of the TEI of Patras and an adjunct lecturer (407/80) in the Department of Computer Engineering and Informatics of the University of Patras, Greece.