Predicting Human miRNA Target Genes Using a Novel Evolutionary Methodology

Korfiati Aigli[1], Kleftogiannis Dimitris[2], Theofilatos Konstantinos[1],
Likothanassis Spiros[1], Tsakalidis Athanasios[1], and Mavroudi Seferina[1]

[1] Department of Computer Engineering and Informatics, University of Patras, Greece
{korfiati,theofilk,likothan,mavroudi}@ceid.upatras.gr,
tsak@cti.gr
[2] Math. and Computer Sciences and Engineering King Abdullah Univ. of Science and Technology
dimitrios.kleftogiannis@kaust.edu.sa

Abstract. The discovery of miRNAs had great impacts on traditional biology. Typically, miRNAs have the potential to bind to the 3'untraslated region (UTR) of their mRNA target genes for cleavage or translational repression. The experimental identification of their targets has many drawbacks including cost, time and low specificity and these are the reasons why many computational approaches have been developed so far. However, existing computational approaches do not include any advanced feature selection technique and they are facing problems concerning their classification performance and their interpretability. In the present paper, we propose a novel hybrid methodology which combines genetic algorithms and support vector machines in order to locate the optimal feature subset while achieving high classification performance. The proposed methodology was compared with two of the most promising existing methodologies in the problem of predicting human miRNA targets. Our approach outperforms existing methodologies in terms of classification performances while selecting a much smaller feature subset.

Keywords: miRNAs, miRNA targets, genetic algorithms, evolutionary computation, Support Vector Machines, Machine Learning classification, multi-objective optimization.

1 Introduction

In recent years the development of high throughput techniques accelerated the discovery of small non-protein-coding regulatory molecules. One type of these are microRNAs (miRNAs). The large family of miRNAs is defined as small (approximately 22 nt) in length, stable molecules which regulate the functions of many other target-genes [1]. Typically, miRNAs have the potential to bind to the 3'untraslated region (UTR) of their mRNA target genes for cleavage or translational repression. The miRNA class is evolutionary conserved and miRNAs have been discovered in animals, flies, plants and viruses [2]. Their targets range from signaling proteins, metabolic enzymes, transcription factors and so on.

I. Maglogiannis, V. Plagianakos, and I. Vlahavas (Eds.): SETN 2012, LNAI 7297, pp. 291–298, 2012.

The very first miRNAs and their targets were discovered experimentally through classical genetic techniques. A description of the experimental techniques and the detailed history of the miRNA genes discovery can be found in [3]. However the experimental identification of miRNA genes and their targets has many drawbacks; cost, time, low specificity are the main technical hurdles.

In order to overcome these limitations and achieve high classification performance many computational approaches have been proposed. The computational methods have proven to be invaluable tools in understanding the biology of miRNAs. Many review papers have already reported the principles of miRNA genes and targets identification and discussed the computational methods that have been applied [4], [5]. Sequence complementarity, thermodynamic stability calculations and evolutionary conservation among species are mainly the most characteristic features which are used to determine the existence of a productive miRNA-mRNA duplex formation [1, 2]. The usage of sequence conservation reduces false positive predictions but some less conserved target sites may be missed leading to low sensitivity. On the other hand, when seed region conservation is not used, the predictors are prone to finding a very large number of predictions and thus to present a low specificity.

At present, TargetScanS [6], PicTar [7], miRanda [8], DIANA-microT [9], miTarget [10] and NBmiRTar [11] are considered to be the most prevailing algorithmic approaches for the prediction of miRNA targets. Although encouraging results are obtained, still existing computational methodologies for the prediction of miRNA targets mainly fail to handle the trade-offs between sensitivity-specificity and interpretability-prediction performance. Furthermore, they select the features which are going to be used as inputs for the prediction either empirically or using simple filtering methods which are incapable of taking advantage of the mutual information between features and of their rate of ability to link with a specific classifier.

In the present work we present a novel computational machine learning approach for the prediction of miRNA targets, which combines the efficiency of Support Vector Machines with Genetic Algorithms for parameters optimization and feature selection. The proposed methodology presents high classification performance combined with a selection of a much smaller feature subset. Thus, it manages to overcome the main problems of existing computational methodologies for the prediction of miRNA targets.

The rest of the article is organized as follows: section 2 describes the implementation of our method. Section 3 provides the experimental results and section 4 concludes the paper.

2 Methodology

2.1 Dataset

In order to distinguish between real and pseudo targets, our model was trained and tested using a relevant biological data set consisting of both positive and negative examples. To ensure the quality of the training data, experimentally verified microRNAs and their targets were collected from the literature. During the data collection step, sequences which were not verified by wet lab experiments were excluded. Targets whose precise binding sites could not be accurately verified were also excluded. The major criterion for including a target into the dataset was the exact binding site of the miRNA-mRNA duplex to be known.

The human microRNAs were downloaded from miRBase database release 18 [12]. The experimentally verified human microRNA targets were downloaded from TarBase version 5c [13] and miRecords release November 2010 [14] databases. After filtering the target binding sites from these sources, the final dataset consisted of 182 human records of miRNAs and their target mRNA binding sites. From this set, 178 records were positive examples, i.e. true targets and the remaining 4 were negative examples, i.e. pseudo targets. Since the current set of 4 confirmed negative examples was insufficient, additional artificial negative examples were generated following the procedure described in [15].

In specific, 3000 artificial mature miRNAs (30 nt long) were used to produce artificial negative examples. Then, MiRanda [8] is used to generate target predictions for these artificial miRNAs, the target results are assumed to be false positive predictions since the query search did not include true miRNAs. The minimum free energy (MFE) and the miRanda score threshold (SC) (in our case set at 25 kcal/mol and 180, respectively) are two important parameters that should be set to increase the stringency of the predictions and thus decrease the selection of weaker false positive predictions [16]. Since using all 3000 artificial miRNAs yielded a large and unmanageable set of predictions, 100 of them were chosen at random to re-query miRanda and from the whole set of false targets 174 were chosen at random to serve as our negative examples. This method is supposed to be superior to choosing miRNA-mRNA duplexes at random. This is because the artificial negative examples we generated resemble more to true duplexes and will therefore lye, with a good probability, closer to the decision boundary.

Between positive and negative examples a 1:1 proportion was used to guarantee that our classifiers will be trained with equal numbers of positive and negative examples. Finally, in order to create the training and the test sets the whole dataset was divided in half at a random manner.

2.2 Feature Set

In order to train our model, we searched for representative features capable of distinguishing efficiently between real and pseudo microRNA targets. As a result we collected most of the features presented in the literature since our method would extract the optimal subset of them. They can be broadly categorized in structural, thermodynamic, positional and ‘motif’ [11] based features. Evolutionary conservation was not used in our approach since it would bias the results to evolutionary conserved miRNA and targets. A total number of 124 features were computed and for the computation, RNAfold program from the Vienna RNA package [17] was used as well as scripts written by us.

2.3 Proposed miRNA Target Prediction Method

A hybrid methodology, implemented in Matlab, was designed and developed for the prediction of miRNA target genes, combining Genetic Algorithms (GA) and Support Vector Machine (SVM) classifiers. The SVM algorithm [18] is the most popular kernel based method and it is considered as a state-of-the-art classification technique able to provide accurate classification models with high generalization ability. Genetic Algorithms [19] are search algorithms inspired by the principle of natural selection. They are useful and efficient if the search space is big and complicated or there is not any available mathematical formulation of the problem. It has been shown that GAs can deal with large search spaces and do not get trapped in local optimal solutions like other optimization algorithms [19].

In our hybrid method the genetic algorithm is used to locate the optimal feature subset and on the same time to tune the *parameters* of the Radial Basis Kernel (C and γ) [20] of the SVM classifier. The produced evolutionary hybrid algorithm mainly consists of the iterative application of the evaluation, selection, crossover and mutation steps in a population of candidate solutions (chromosomes) which are initially randomly generated. Each chromosome is consisted of *feature genes* that encode the best feature subset and *parameter genes* that encode the best choice of parameters (binary encoding was used). The proposed method is depicted in detail in Fig.1.

The size of the initial population was set to 20 chromosomes and the termination criterion stops the evolution when the population is deemed as converged. Specifically, the population is deemed as converged when the average fitness across the current population is less than 5% away from its best fitness. Alternatively, the algorithm stops when the maximum number of 100 generations is reached. The proposed fitness function which was used to evaluate each candidate solution is the one described in equation (1):

$$Fitness = 0.5 \cdot Accuracy + 0.5 \cdot GeometricMean - 0.001 \cdot Selected_{Features} - 0.0001 \cdot SupportVectors \tag{1}$$

where Accuracy is the SVM's accuracy, GeometricMean is the geometric mean of sensitivity and specificity, Selected_Features is the selected feature subset size and SupportVectors is the number of support vectors included in the trained SVM model. This fitness function is used in order to balance classification performance, sensitivity-specificity tradeoff, the complexity of the feature set (which relates to the interpretability of the model) and the classification model's complexity (and thus its generalization ability). The values for the constant multipliers in our multiobjective fitness function were set as shown in equation (1) to weight our goals in the following order of significance (from the least significant to the most significant):

Simplicity of Classification model< Complexity of feature set <Sensitivity-specificity tradeoff = Classification performance

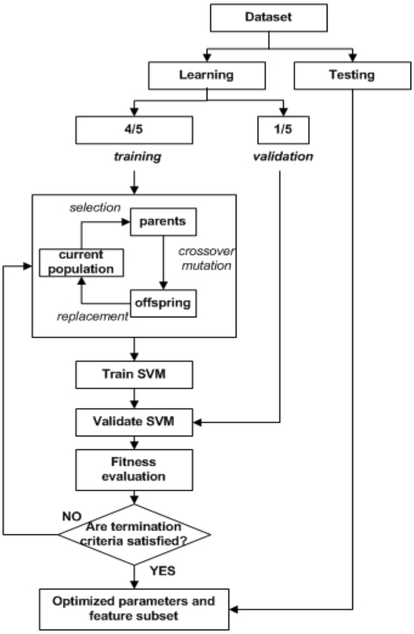

Fig. 1. Flow Chart of the proposed method

The selection scheme which was applied in the proposed hybrid methodology was Rank based roulette wheel selection [21] in order to control the pressure of natural selection. Following, this selection scheme the proposed evolutionary algorithm forces our population to areas of better solutions while minimizing the probability of reducing the population's diversity and thus minimizing the possibility of getting trapped in local optimal.

The two main genetic operators of a Genetic Algorithm are crossover and mutation. For the crossover operator, two-point crossover was used to create two offsprings from every two selected parents. The parents are selected at random, two crossover points are selected at random and two offsprings are made by exchanging genetic material between the two crossover points of the two parents. The crossover probability was set equal to 0.9 to leave some part of the population to survive unchanged to the next generation.

Most studies on the selection of the optimal mutation rate parameter coincide that a time-variable mutation rate scheme is usually preferable than a fixed mutation rate [22]. Accordingly, we propose the dynamic control of the mutation parameter using equation (2):

$$Pm(n) = 0.2 - n \cdot \frac{0.2 - \frac{1}{P_S}}{\mathbf{MAX_G}} \tag{2}$$

where n is the current generation, PS is the size of the population and MAXG is the maximum generation specified by the termination criteria. Using equation (2), we start with a high mutation rate for the first generations and then gradually decrease it over the number of generations. In this way global search characteristics are adopted in the beginning and are gradually switched to local search characteristics for the final iterations. The mutation rate is reduced with a smaller step when a small population size is used in order to avoid stagnation. For bigger population sizes the mutation rate is reduced with a larger step size since a quicker convergence to the global optimum is expected.

3 Results

In order to evaluate the performance of our methodology, we compared its test set performance with the corresponding performance, of the miTarget[10] and NBmiRTar[11] classifiers. For the constructed data set, due to the stochastic nature of the proposed methodology and the randomization process for the splitting of training and test sets in every execution, we ran the experiments 20 times and computed the average values. Table 1 summarizes the results achieved.

As we can easily observe in Table 1 our method achieves a significantly better classification performance and at the same time it uses a smaller feature set.

Table 1. Comparative classification performance

Method	Features	Accuracy	Specificity	Sensitivity
miTarget	41	93,93%	89,77%	89,77%
NBmiRTar	67	85,73%	80,13%	91,48%
Proposed Methodology	**38,2**	**99,10%**	**98,24%**	**100,00%**

With a further examination of the frequency of appearance of the extracted features, we observed that the following subset of 5 features had a frequency over 80%:

- number of matches in seed part
- total number of AU matches
- number of bulges of length 5 in out-seed part
- free energy of the out-seed part
- position 3 with a GC match, an AU match, a GU match or a mismatch

These results confirm our hypothesis that the suggested methodology is capable of identifying representative features which optimize the performance. It is also observed that this extracted feature subset contains various types of information, including structural, thermodynamic and positional features. The results also indicate how important for the prediction the out-seed region is, since most of the extracted features concern this region.

4 Conclusion and Future Work

In the present paper we have introduced a novel methodology for the prediction of miRNA targets. This methodology consists of a hybrid combination of a modified genetic algorithm with an SVM classifier using a novel multi-objective fitness function. In contrast to previous approaches where the performance of the classifiers strongly depends on the careful (mostly by hand) pre-selection of the optimal features, our algorithm accepts all available features as input (without any restrictions concerning independency) and automatically generates a small optimal feature subset. At the same time it finds the optimal SVM parameters C and γ for the optimal feature set and produces prediction models of high classification performance.

The proposed methodology was compared with two of the latest existing methodologies and outperformed them in both classification performances and number of the selected features. Using methodologies like the proposed one, that can find a small informative subset of features which does not include features with mutual information or irrelevant features, may help biologists to gain a better understanding of the miRNA targeting procedure.

In order to further enhance the interpretability of our classifier, as future work, we will extract fuzzy rules from the SVM classifier using the methodology described in [23]. Finally, our future plans involve the development of a web-based tool for the online prediction of miRNA targets based on the proposed methodology.

Acknowledgments. (This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

References

1. Bartel, D.P.: MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116(2), 281–297 (2004)
2. Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A., Tuschl, T.: New microRNAs from mouse and human. RNA 9(2), 175–179 (2003)
3. Lai, E.C.: microRNAs: runts of the genome assert themselves. Curr. Biol. 13(23), R925–R936 (2003)
4. Mendes, N.D., Freitas, A.T., Sagot, M.-F.: Current tools for the identification of miRNA genes and their targets. Nucleic Acids Res. 37(8), 2419–2433 (2009)
5. Li, L., Xu, J., Yang, D., Tan, X., Wang, H.: Computational approaches for microRNA studies: a review. Mamm. Genome 21(1-2), 1–12 (2010)
6. Lewis, B.P., Burge, C.B.: Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120(1), 15–20 (2005)
7. Grun, D., Wang, Y.L., Langenberger, D., Gunsalus, K.C., Rajewsky, N.: MicroRNA target predictions across seven Drosophila species and comparison to mammalian targets. PLoS Comput. Biol. 1(1), 51–66 (2005)
8. Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., Marks, D.S.: MicroRNA targets in Drosophila. Genome Biol. 5(1), R1.1–R1.14 (2005)
9. Kiriakidou, M., Nelson, P.T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z., Hatzigeorgiou, A.: A combined computational- experimental approach predicts human microRNA targets. Genes Dev. 18, 1165–1178 (2004)

10. Kim, S.K., Nam, J.W., Rhee, J.K., Lee, W.J., Zhang, B.T.: miTarget: microRNA target-gene prediction using a support vector machine. BMC Bioinformatics 7, 411–422 (2006)
11. Malik, Y., Jung, S., Kossenkov, A., Showe, L., Showe, M.: Naïve Bayes for microRNA target predictions—machine learning for microRNA targets. Bioinformatics 23(22), 2987–2992 (2007)
12. Griffiths-Jones, S.: The microRNA Registry. Nucl. Acids Res. 32(suppl. 1), D109–D111 (2004)
13. Papadopoulos, G.L., Reczko, M., Simossis, V.A., Sethupathy, P., Hatzigeorgiou, A.G.: The database of experimentally supported targets: a functional update of TarBase. Nucleic Acids Res. 37, D155–D158 (2009)
14. Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., Li, T.: miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res. 37, D105–D110 (2009)
15. Saetrom, O., Snøve, O., Saetrom, P.: Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. RNA 11, 995–1003 (2005)
16. Hsu, P.W.: miRNAMAP: genomic maps of microRNA genes and their target genes in mammalian genomes. Nucleic Acids Res. 34, D135–D139 (2006)
17. Hofacker, I.L.: Vienna RNA secondary structure server. Nucleic Acids Res. 31(13), 3429–3431 (2003)
18. Lewis, D.P., Jebara, T., Noble, W.S.: Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure. Bioinformatics 22, 2753–2760 (2006)
19. Holland, J.: Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT Press, Cambridge (1995)
20. Vapnik, V.N.: The nature of statistical learning theory. Springer (2000)
21. Jadaan, O., Rao, C.R., Rajamani, L.: Parametric Study to Enhance Genetic Algorithm Performance, Using Ranked based Roulette Wheel Selection method. In: InSciT 2006, Merida, Spain, vol. 2, pp. 274–278 (2006)
22. Thierens, D.: Adaptive Mutation Rate Control Schemes in Genetic Algorithms. In: Proceedings of the 2002 IEEE World Congress on Computational Intelligence: Congress on Evolutionary Computation, pp. 980–985 (2002)
23. Mavroudi, S., Katsanos, P., Papadimitriou, S., Likothanassis, S.: Transparent Classification Process of Bioinformatics Data with an Approximated Support Vector Fuzzy Inference System. In: The International Special Topic Conference on Information Technology in Biomedicine (ITAB 2006), Ioannina, Epirus Greece, October 26-28 (2006)