Data Mining for MicroRNA Gene Prediction

On the Impact of Class Imbalance and Feature Number for MicroRNA Gene Prediction

Müşerref Duygu Saçar
Molecular Biology and Genetics
Izmir Institute of Technology
Gulbahce, Urla, Izmir, Turkey
duygusacar@gmail.com

Jens Allmer
Molecular Biology and Genetics
Izmir Institute of Technology
Gulbahce, Urla, Izmir, Turkey
jens@allmer.de

***Abstract*—MicroRNAs (miRNAs) are small, non-coding RNAs which are involved in the posttranscriptional modulation of gene expression. Their short (18-24) single stranded mature sequences are involved in targeting specific genes. It turns out that experimental methods are limited and that it is difficult, if not impossible, to establish all miRNAs and their targets experimentally. Therefore, many tools for the prediction of miRNA genes and miRNA targets have been proposed. Most of these tools are based on machine learning methods and within that area mostly two-class classification is employed. Unfortunately, truly negative data is impossible to attain and only approximations of negative data are currently available. Also, we recently showed that the available positive data is not flawless. Here we investigate the impact of class imbalance on the learner accuracy and find that there is a difference of up to 50% between the best and worst precision and recall values. In addition, we looked at increasing number of features and found a curve maximizing at 0.97 recall and 0.91 precision with quickly decaying performance after inclusion of more than 100 features.**

Keywords—microRNA; machine learning; data mining; class imbalance; feature selection; miRNA gene prediction

I. Introduction

MicroRNAs (miRNAs) were discovered about two decades ago [1] and have since attracted growing interest. MicroRNAs are best characterized by their canonical pathway. First an enzyme called microprocessor cleaves a hairpin like structure from a nascent RNA (pri-miRNA) and thus a) effectively cleaves the RNA into three smaller pieces and b) produces a pre-miRNA. This pre-miRNA has a stem-loop structure which resembles a hairpin. Upon its production, Exportin-5 channels the hairpin into the cytosol where the loop is cleaved off to leave a short double stranded RNA of between 18 to 24 nucleotides in length. One of the strands is then incorporated into the RNA induced silencing complex (RISC). The incorporated strand serves as a key to target mRNAs by its complement.

MicroRNAs originate from anywhere in a genome [2] and may target any gene with a complementary sequence within its mRNA. They have been found in most taxa and viruses have even been shown to regulate host encoded genes [3] but miRNAs are tightly regulated and often confined to a specific tissue, developmental stage, or stress response [4].

MicroRNAs can be discovered by experimental methods like directional cloning of endogenous small RNAs, but such methods are time consuming or expensive [5]. Together with their controlled timely and locational expression and the additional constrain that both miRNA and its target must be co-expressed, it becomes obvious that trying to experimentally determine all miRNAs, their targets, and their interactions is futile with current technology.

This has led to the proposition of many miRNA detection algorithms [6]. Many of these algorithms are using machine learning for the detection of miRNAs [6], [7]. Support vector machine classification, using positive and negative examples for training and testing of the classifier, dominates the field (Saçar and Allmer, Methods in Molecular Biology, 2013, in press). The dependency on positive and negative data is problematic since, as outlined above, it will not be possible to create truly negative data using experimental strategies for any eukaryotic organism. This entails that the negative class is likely contaminated with many false negative examples which affects classification accuracy. Thus establishment of negative data is difficult [8–10]. So far, only one one-class classifier has been proposed for *ab inito* miRNA detection to overcome the limitation [10].

Unfortunately, the establishment of positive data is also not as straightforward as it could be expected. We recently showed that miRBase [11], the largest database for miRNAs, contains dubious examples for human and that, if all positive human miRNAs are used for classification, the accuracy is less than if the more stringently annotated examples from miRTarBase [12] are being used [13].

Here we investigate the impact of largely differing numbers of positive and negative examples on machine learning for *ab initio* miRNA prediction. This class imbalance problem is especially pronounced for miRNAs since it has been estimated that there are millions of hairpins in the human genome [14], but less than 2000 true miRNAs. Some larger negative datasets have been proposed and we base our analysis on these along with positive examples from miRBase. Using different combinations of positive and negative examples of increasing size, we here show that classification accuracy is not a good measure and that precision (up to ~30%) and recall (up to 50%) are strongly affected. In addition, we used

This work was in part supported by an award for outstanding young scientist by the Turkish Academy of Sciences (TÜBA) awarded to Jens Allmer.

978-1-4799-0701-4/13/$31.00 ©2013 IEEE

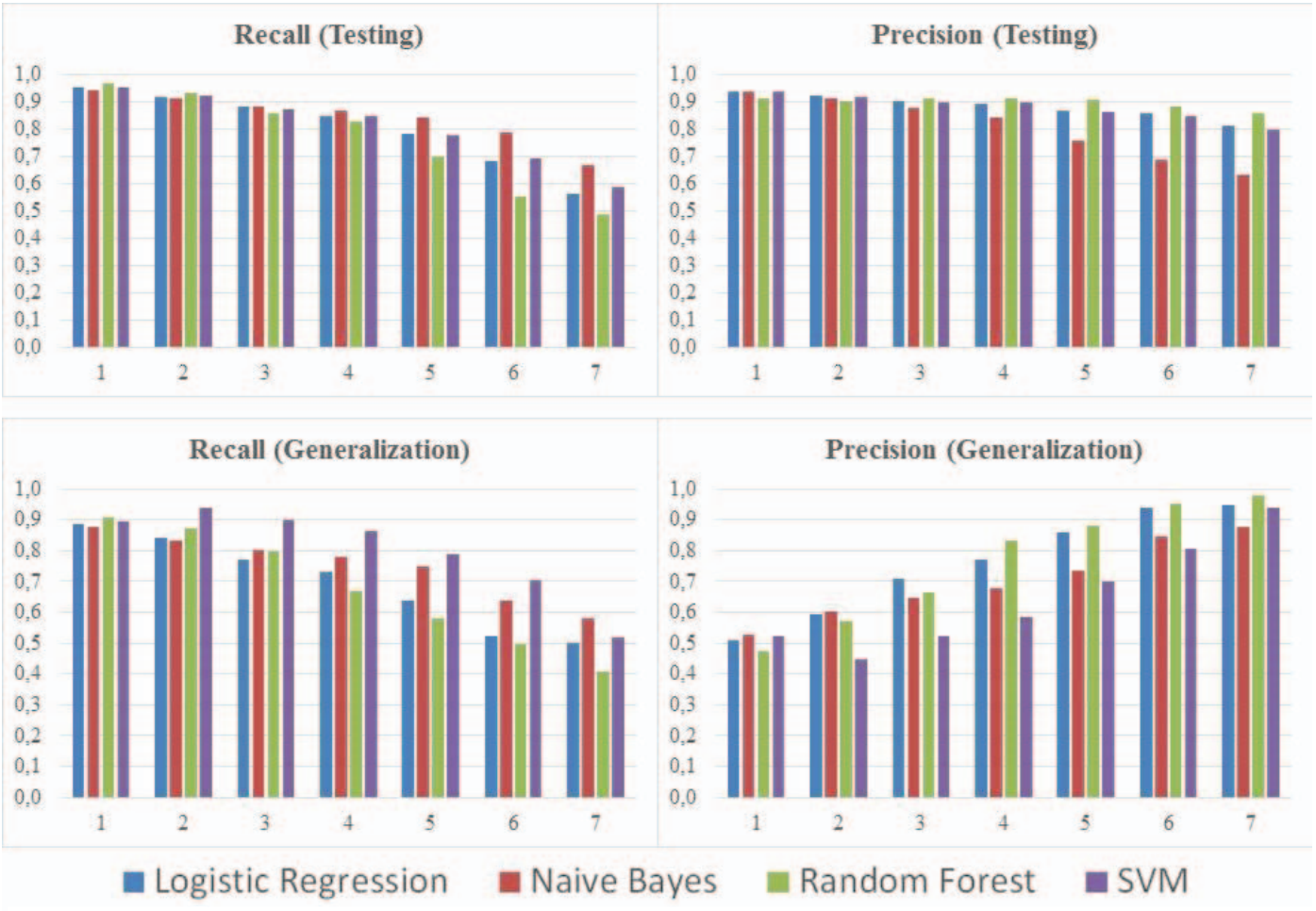

Fig. 1. Recall and precision after training on the test dataset using 10 fold cross validation for four different classifiers (top left and right). Recall and precision for the best classifier tested on new examples that were not part of training or test set (bottom left and right). Numbers 1 through 7 correspond to different composition of positive and negative data during training. 1:1600 positive and 800 negative examples; 2 1600, 1600; 3: 1600-3200; 4: 1600-5000; 5: 1600-10000; 6: 1600-25000; 7: 1600-50000.

the trained classifier for generalization and the effect of class imbalance, during training, on generalization is just as devastating as in the testing phase. We further investigated the impact of increasing the number of features for the best model (1600 positive and 1600 negative examples). This analysis shows that with an increase in features the returns are diminishing and that with the feature set we have (about 350 features) an upper bond of slightly more than 0.9 can be achieved for recall and precision.

II. Materials and Methdos

A. Data sets

Positive examples for human miRNAs were obtained from miRBase (http://www.mirbase.org/). To be able to examine class imbalance effect, the largest available negative dataset for human was used (http://adaa.polsl.pl/agudys/huntmi/huntmi.htm) [15]. Of this dataset about 50000 sequences were used for testing. For generalization the remaining about 18000 examples from the dataset were used. In addition to that, miRNA examples from ENSEMBLE (≈3200) [16], the complete pseudo data set (≈9000) as described in [5], and random sequences with the same length range with human miRNAs (1400) were used. The generalization dataset consists of approximately 32000 examples with most of them being negative.

B. Features

Approximately 12 studies performed *ab initio* miRNA prediction. We implemented all features suggested in four selected *ab initio* miRNA detection studies [5], [8], [17], [18]. In addition to implementing these features, we generalized the features and normalized them to miRNA stem length or miRNA hairpin length where appropriate. The total number of features that we analyzed amounts to about 330.

C. Data Mining

Orange Canvas (http://orange.biolab.si/) [19], a widely used open source data visualization and analysis tool was used for data mining. In order to create data sets in differing sizes,

negative data was randomly sampled without replacement. At the end 7 different data sets with the most commonly used 10 features found in 12 *ab initio* miRNA gene prediction studies (hll, bpp/hpl, hpmfe_rf/hpl, hpl, *U(((, #U(((/hpl, *U(.(, #U(./hpl, *C(.(, #C(./hpl, *A..., #A.../hpl, *G(((, #G(((/hpl, Q/hpl) [13] were produced for training. In the first dataset (1) 1600 positive and 800 negative examples were used. Dataset 2 contained 1600-1600; 3: 1600-3200; 4: 1600-5000; 5: 1600-10000; 6: 1600-25000; and dataset 7: 1600-50000 positive and negative examples, respectively. These data sets were used for training four different classifiers (Logistic Regression, Naïve Bayes, Random Forest, SVM) using the default settings in Orange Canvas. The models produced by each classifier for all data sets were saved and these models were later used for generalization on a different data set which does not include the data used in the training stage.

TABLE I. INFORMATION GAIN FOR THE 87 FEATURES WITH HIGHEST GAIN AMONG ALL DEFINED FEATURES.

Attribute	Infor. Gain	Attribute	Infor. Gain	Attribute	Infor. Gain
#C.../sl	1,00	Q	1,00	dscs/nl	0,67
#U.../sl	1,00	#C.(./sl	0,99	efq	0,67
#A..(/sl	1,00	Tm	0,99	ediv	0,66
#A(../sl	1,00	#U.(./sl	0,99	saln/hpl	0,66
#G(((/sl	1,00	#A.(./sl	0,98	bpp/sl	0,66
#A(((/sl	1,00	dH/sl	0,96	mbs/sl	0,66
#U(((/sl	1,00	dS/sl	0,96	mbs/hpl	0,66
#U..(/sl	1,00	dS/hpl	0,95	lsr(%bp)/hpl	0,64
#G.../sl	1,00	dH/hpl	0,95	lsr(%bp)/sl	0,62
#C..(/sl	1,00	Tm/sl	0,94	#nial_h/sl	0,61
#C((./sl	1,00	hpmfe_rf/sl	0,94	#nial_h/hpl	0,61
#A.../sl	1,00	hpmfe_rf_l	0,94	lscm/nl	0,60
#A((./sl	1,00	dG/sl	0,93	adal/hpl	0,59
#U.((/sl	1,00	hpmfe_rf/hpl	0,91	bpp/saln	0,59
#U(../sl	1,00	Q/sl	0,91	#gih/saln	0,58
#C(((/sl	1,00	dG/hpl	0,89	#goh/saln	0,58
#U((./sl	1,00	efe	0,88	asal/hpl	0,58
#G(../sl	1,00	Tm/hpl	0,88	nl/sl	0,57
#C(.(/sl	1,00	bpd/sl	0,86	nl/hpl	0,56
#C(../sl	1,00	Q/hpl	0,81	lscm/hpl	0,54
#C.((/sl	1,00	bpd/hpl	0,74	st(A-U)/hpl	0,52
#G((./sl	1,00	st(G-C)/hpl	0,72	mwmF/hpl	0,51
#A.((/sl	1,00	l(lsr)/hpl	0,70	st(A-U)/sl	0,50
#G.((/sl	1,00	st(G-C)/sl	0,69	*G(((	0,48
#G(.(/sl	1,00	bpp/hpl	0,69	*A...	0,47
#U(.(/sl	1,00	mwm/sl	0,69	#A++#U/hpl	0,47
#G..(/sl	1,00	bpp/nl	0,68	#U++#A/hpl	0,47
#A(.(/sl	1,00	hpmfe_rf	0,68	%U++%A	0,47
#G.(./sl	1,00	l(lsr)/sl	0,68	%A++%U	0,47

In order to test the influence of the number of features used on precision and recall of classification, the dataset with 1600 positive and 1600 negative examples was selected and classifiers were trained with varying amount of features. Firstly the features are ranked based on the information gain score (see Table I for a sample) and by starting from the 1st feature to the last in the list, 8 data files were produced with 5, 10, 20, 30, 50, 100, 200, and 334 features. Additionally, by starting from the feature with the least information gain to the 1st one in the list, 8 data files were produced with 5, 10, 20, 30, 50, 100, 200, and 334 features. These 16 different data sets were used for classification using Naive Bayes since it was more robust for larger number of features than other algorithms in Orange Canvas.

III. Results and Discussion

A. Class Imbalance Effect

Previously, we compared 4 different *ab inito* miRNA gene prediction studies [7] and found that their performance is quite different from their published performance when compared on the same dataset. We then became interested in analyzing the effect of class imbalance on the classification accuracy on test data and on the generalization of the trained classifier. We prepared a number of scenarios with a fixed number of 1600 positive examples and the number of negative examples ranging from 800 to 50000.

For all cases we assessed the generalization performance of the best trained classifier on ~30000 previously unused examples. During training and testing there is a decrease in both precision and recall with increasing class imbalance (Fig. 1). Recall is more affect (up to ~50%) while precision remains largely constant across the test cases for most classifiers except for Naïve Bayes which drops by about 30%. Applying the trained classifiers to the unseen data leads to a drop in precision and recall for most cases compared to the performance during training/testing. It is apparent from Fig. 1 that during generalization (bottom panes) recall drops significantly (up to ~50%) while precision increases at the same time (up to ~50%). This suggests that accuracy may not be a good measure to report for the performance of miRNA gene prediction based on machine learning and that it is prerogative to report precision and recall or sensitivity and specificity, instead.

There is slight variation among the precision and recall values for the four classifiers tried in this study but their relative performance is quite similar. Therefore, we cannot suggest using a particular classification method over any of the other methods. We used the default settings for all classifiers, but we believe that a slight increase in performance could be achieved by optimizing the parameters which was not a subject of this study. Furthermore, since there is some deviation among classification methods, an in chorus approach of using multiple classifiers to detect miRNAs may be useful.

B. Effect of Feature Number

In a recent study, we assessed the performance of four *ab initio* miRNA gene prediction tools and found that a large

number of features does not necessarily lead to better performance in respect to recall and precision [7]. Before that we assessed potential features and concluded that not the number of features but the ensemble prediction of miRNA genes and targets adds to the specificity of miRNA gene prediction [20].

Since we implemented a large number of features, we were interested in how increasing number of features on the best training dataset from *A*. affects classification performance. We ranked all features according to information gain (Table I) and then started with the 5 most extreme features and included more features until all features were included in the classification. This leads to two curves for precision and recall (Fig. 2) one set which starts from low performance (initialized from the bottom of the ranked feature list; green and purple) and one which starts with high performance (initialized from the top of the ranked feature list; blue and red). When features are selected from the bottom first, the trained classifiers never reach to recall or precision of classifiers trained starting with features with the largest information gain. With increasing number of features, the precision and recall values of the trained classifiers fall into the same range. They cannot be expected to reach to exactly the same values as an element of randomness is introduced through the use of 10 fold cross validation.

We included the assessment of adding features from the bottom of the list to highlight that there is a significant impact of feature selection on precision and recall. With the ability to compare the two sets it is more convincing to see a clear drop in performance between 100 and 200 features and attribute it to the information content of the features. There is a slight increase in precision up to 100 features while recall stays quite constant for the selection of features from top of the list.

From this analysis we suggest that in future studies no more than 100 of these features should be used. In addition to this analysis we looked at the correlation among the features. From Fig. 3, which only displays the correlation of a subset of the features with highest information gain, it can be deduced that many of the features are highly correlated. Table I further supports this notion of feature correlation and it can be seen that many similar features are among the best ones, displayed in Fig. 1. Clearly, this is not desirable for machine learning since it may lead to the over representation of a particular feature on the learned model. Therefore, future studies need to ensure that features should have low to no correlation.

IV. Conclusion

Class imbalance, coupled with the fact that both positive and negative miRNA examples are questionable, has a great impact on the training and the generalization ability of classifiers aimed to detect miRNA genes. We here show that

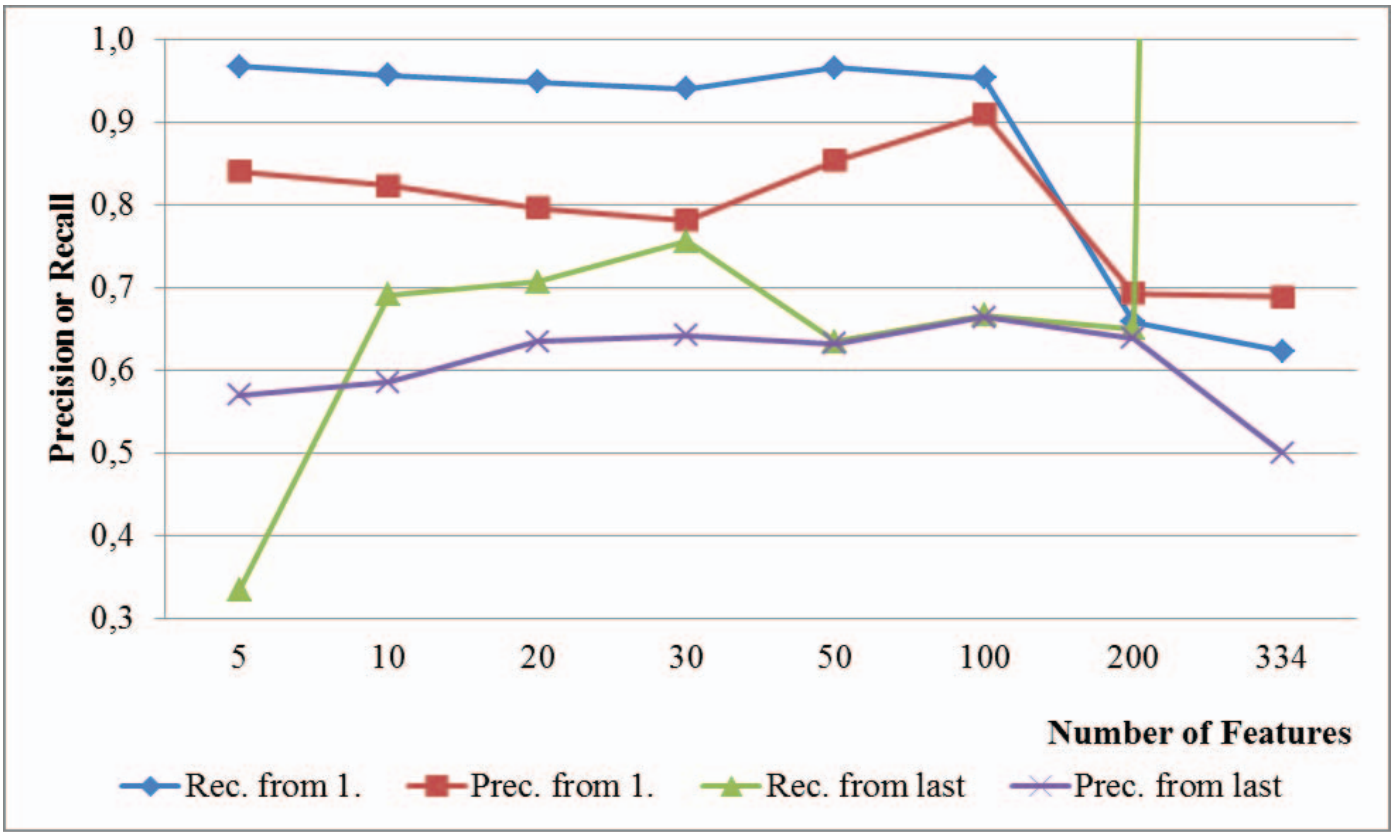

Fig. 2. Rec. from 1. refers to recall for classifiers (here Naïve Bayes was used) trained with increasing number of features selected from the top of the list of ranked features according to information gain. Prec. from 1. refers to the precision of classifiers trained in that manner. Rec. and Prec. from last refer to the recall and precision of classifiers with feature number increasing from the bottom of the ranked feature list. The calculation of the recall for 334 features from last was unsuccessful and therefore the value is out of bounds and not displayed in the figure.

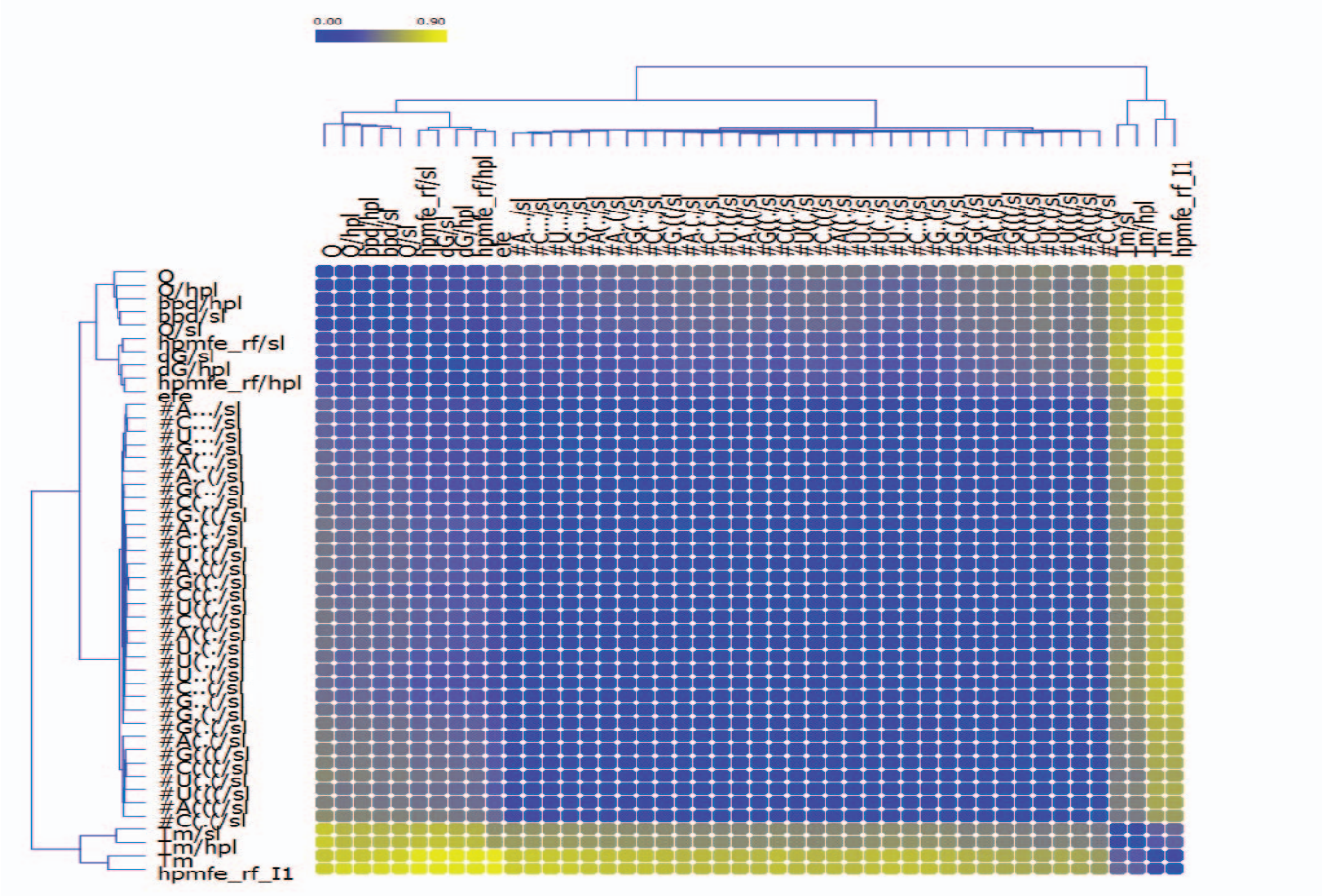

Fig. 3. Correlation among the features with the largest information gain (see Table I). Blue indicates high and yellow low correlation among the features. Features are further clustered hierarchially across the top and on the left hand side.

for currently available data it is best to use 1600 positive and 1600 negative examples. We are not able to suggest the use of a particular machine learning method to be most successful with miRNA detection, but we believe that parallel usage of multiple classifiers will be most successful in the future.

Feature selection is an NP-hard process [21] but the number of features proposed, including their derivatives, in *ab initio* miRNA gene prediction is much larger than 300 (we are in the process of implementing all features that have been proposed) and the lower bound of the number of features that are needed for proper feature selection is unknown. Here we show that the upper bound of the subset of features we implemented seems to be 100 but it would be of use to implement and normalize the remaining features that have been proposed. Selecting the best set of 100 features from 300 features, however, is beyond our computational abilities. There are heuristic approaches and we are planning to try them in the future.

Another issue is correlation among features (Fig. 3). This problem may reduce the feature selection problem since it may aid in reducing the number of features that can be selected from, if solved. We here used attribute correlation as implemented in Orange Canvas, but it turns out that some attributes which are logically/biologically strongly correlated were not reported to be correlated by the given method. For example, one of the features of attributes like the p-value of the minimum free energy and the minimum free energy itself (data not shown), which are probably not reported to be correlated since the range of their values is largely different, should be removed from the features. The same is true for many of the features and their normalized versions. Therefore, we aim to manually inspect all attributes in the future and first remove all obviously correlated features before we again try attribute correlation or any other feature selection algorithm.

References

[1] R. C. Lee, R. L. Feinbaum, and V. Ambros, "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.," *Cell*, vol. 75, no. 5, pp. 843–54, Dec. 1993.

[2] A. Rodriguez, S. Griffiths-Jones, J. L. Ashurst, and A. Bradley, "Identification of mammalian microRNA host genes and transcription units," *Genome Research*, vol. 14, no. 10A, pp. 1902–1910, Oct. 2004.

[3] S. Pfeffer, M. Zavolan, F. A. Grässer, M. Chien, J. J. Russo, J. Ju, B. John, A. J. Enright, D. Marks, C. Sander, and T. Tuschl, "Identification of virus-encoded microRNAs," *Science*, vol. 304, no. 5671, pp. 734–6, Apr. 2004.

[4] A. Aravin and T. Tuschl, "Identification and characterization of small RNAs involved in RNA silencing.," *FEBS Letters*, vol. 579, no. 26, pp. 5830–40, Oct. 2005.

[5] 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. 1321–30, Jun. 2007.

[6] J. Allmer and M. Yousef, "Computational methods for ab initio detection of microRNAs.," *Frontiers in genetics*, vol. 3, p. 209, Jan. 2012.

[7] M. D. Saçar and J. Allmer, "Comparison of four Ab Initio MicroRNA Prediction Tools," in *4th International Conference on Bioinformatics Models, Methods and Algorithms*, 2013.

[8] J. Ding, S. Zhou, and J. Guan, "MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features.," *BMC bioinformatics*, vol. 11 Suppl 1, no. Suppl 11, p. S11, Jan. 2010.

[9] W. Ritchie, D. Gao, and J. E. J. Rasko, "Defining and providing robust controls for microRNA prediction.," *Bioinformatics (Oxford, England)*, vol. 28, no. 8, pp. 1058–61, Apr. 2012.

[10] M. Yousef, S. Jung, L. C. Showe, and M. K. Showe, "Learning from positive examples when the negative class is undetermined--microRNA gene identification.," *Algorithms for molecular biology*, vol. 3, p. 2, Jan. 2008.

[11] S. Griffiths-Jones, "miRBase: microRNA sequences and annotation.," *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, vol. Chapter 12, p. Unit 12.9.1–10, Mar. 2010.

[12] S.-D. Hsu, F.-M. Lin, W.-Y. Wu, C. Liang, W.-C. Huang, W.-L. Chan, W.-T. Tsai, G.-Z. Chen, C.-J. Lee, C.-M. Chiu, C.-H. Chien, M.-C. Wu, C.-Y. Huang, A.-P. Tsou, and H.-D. Huang, "miRTarBase: a database curates experimentally validated microRNA-target interactions.," *Nucleic acids research*, vol. 39, no. Database issue, pp. D163–9, Jan. 2011.

[13] M. D. Saçar, H. Hamzeiy, and J. Allmer, "Can MiRBase Provide Positive Data for Machine Learning for the Detection of MiRNA Hairpins?," *Journal of integrative bioinformatics*, vol. 10, no. 2, p. 215, Jan. 2013.

[14] I. Bentwich, "Prediction and validation of microRNAs and their targets," *FEBS Letters*, vol. 579, no. 26, pp. 5904–5910, Oct. 2005.

[15] A. Gudy, M. W. Szcze Niak, M. Sikora, and I. Makalowska, "HuntMi: an efficient and taxon-specific approach in pre-miRNA identification.," *BMC bioinformatics*, vol. 14, no. 1, p. 83, Mar. 2013.

[16] P. Flicek, I. Ahmed, M. R. Amode, D. Barrell, K. Beal, S. Brent, et al., "Ensembl 2013.," *Nucleic acids research*, Nov. 2012.

[17] P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun, and Z. Lu, "MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features," *Nucleic Acids Research*, vol. 35, no. Web Server issue, pp. W339–344, Jul. 2007.

[18] I. Bentwich, "Identifying human microRNAs.," *Current Topics In Microbiology And Immunology*, vol. 320, pp. 257–69, Jan. 2008.

[19] T. Curk, J. Demsar, Q. Xu, G. Leban, U. Petrovic, I. Bratko, G. Shaulsky, and B. Zupan, "Microarray data mining with visual programming.," *Bioinformatics*, vol. 21, no. 3, pp. 396–8, 2005.

[20] M. V. Cakir and J. Allmer, "Systematic computational analysis of potential RNAi regulation in Toxoplasma gondii," in *Health Informatics and Bioinformatics (HIBIT), 2010 5th International Symposium on*, 2010, pp. 31–38.

[21] E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, vol. 209, no. 1–2, pp. 237–260, 1998.