ncRNA-class Web Tool: non-coding RNA Feature Extraction and pre-miRNA classification Web Tool

Dimitrios Kleftogiannis[1] , Konstantinos Theofilatos[2], Stergios Papadimitriou[4], Athanasios Tsakalidis[2], Spiros Likothanassis[2], Seferina Mavroudi[2,3]

[1] Math. & Computer Sciences & Engineering King Abdullah Univ. of Science and Technology, Saudi Arabia
[2] Department of Computer Engineering and Informatics, University of Patras, Greece
[3] Department of Social Work, School of Sciences of Health and Care, Technological Educational Institute of Patras
[4] Department of Information Management, Technological Institute of Kavala, Greece
{theofilk, tsak, likothan, mavroudi } @ceid.upatras.gr ,
dimitrios.kleftogiannis@kaust.edu.sa, sterg@teikav.edu.gr

Abstract. Until recently, it was commonly accepted that most genetic information is transacted by proteins. Recent evidence suggests that the majority of the genomes of mammals and other complex organisms are in fact transcribed into non-coding RNAs (ncRNAs), many of which are alternatively spliced and/or processed into smaller products. Non coding RNA genes analysis requires the calculation of several sequential, thermodynamical and structural features. Many independent tools have already been developed for the efficient calculation of such features but to the best of our knowledge there does not exist any integrative approach for this task. The most significant amount of existing work is related to the miRNA class of non-coding RNAs. MicroRNAs (miRNAs) are small non-coding RNAs that play a significant role in gene regulation and their prediction is a challenging bioinformatics problem. Non-coding RNA feature extraction and pre-miRNA classification Web Tool (ncRNA-class Web Tool) is a publicly available web tool (http://150.140.142.24:82/Default.aspx) which provides a user friendly and efficient environment for the effective calculation of a set of 58 sequential, thermodynamical and structural features of non-coding RNAs, plus a tool for the accurate prediction of miRNAs.

Keywords: ncRNAs, miRNA, feature calculation, miRNA prediction, web tool

1 Introduction

The term non coding RNA (ncRNA) refers to the RNA sequences which are transcribed from DNA and are not translated to proteins. Until recently, the most well-known ncRNA was the transfer RNA (tRNA) which is involved in the mRNA translation [1]. The last two decades the rapid development of the biological and biomedical research led to the identification of many more ncRNA classes. The fact that they are not encoded to proteins does not mean that they do not play a significant

role on the underlying biological processes [2]. Striking examples are the ribosomal RNAs (rRNA) which constitute a central component of the ribosome , the small nuclear RNAs (snRNA) which are involved in the splicing processes , the snoRNAs which modify other rRNAs , the bacterial transfer-messenger RNAs (tmRNA), the PIWI interacting RNAs (piRNA) which are related to the DNA methylation and of course the class of the microRNAs (miRNA) . All the previous examples of ncRNAs comprise a hidden regulatory layer with significant effects on the gene expression, the chromatin architecture and on the translational and transcriptional modifications. Usually, the abnormal ncRNA activity and dysregulation is linked to many diseases such as neurological conditions and tumor genesis. Thus, the study of the functionality and the effective identification of these molecules may lead to more sophisticated therapeutic strategies. The last years, the identification of these ncRNAs classes and the determination of their regulatory networks constitute an open research area with a lot of promises. Many review papers address their role in human diseases, the limitations that derive from their study and the great challenges for future research.

The most significant amount of work is related to the miRNA class of non-coding RNAs. 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 [3]. Typically, miRNAs have the potential to bind to the 3'untraslated region (UTR) of their mRNA target genes for cleavage or translational repression. The very first miRNAs and their targets were discovered experimentally through the classical genetic techniques [4]. However, their small size and the fact that many of them are expressed in very low levels make their experimental identification problematic.

In order to overcome the hurdles and the limitations of the experimental identification techniques, many computational approaches have been developed [5]. The earliest approaches for discovering pre-miRNAs are based on comparative techniques and they can identify miRNAs with close homologs among species. In opposite to the comparative methods, the non-comparative ones do not rely on the heavily phylogenetic conservation. They emphasize on machine learning algorithms to scan for miRNA candidates. In every Machine Learning method the first step consists of the computation of topological, sequential, thermodynamical and other characteristics of the miRNAs. Then, a classifier is trained based on these features and a model that is capable of predicting candidate miRNAs is constructed.

Up to our knowledge, the problem of the microRNA prediction is well studied and more than thirty methodologies have been proposed. However, a small portion of them is supplemented by a user friendly and functional web tool that integrates additional services. That fact becomes an obstacle for the future research agenda and slows down the development of prediction tools for the other classes of non-coding RNAs. The main disadvantages of the existing tools are their difficulty to extract a representative feature set, the lack of a tool that enables batch prediction for a big number of candidate sequences, the limited possibilities of the existing standalone applications and finally the absence of a relational database that stores all the processed sequences.

In the present paper we demonstrate that the careful choices in the development of a microRNA genes classifier accompanied with the development of a functional web tool can substantially improve the future research and accelerate the development of

more sophisticated tools for the other classes of non-coding RNAs. The ncRNA-class Web Tool is introduced as a user friendly web tool which provides the effective calculation of a set of 58 sequential, thermodynamical and structural features of non-coding RNAs, and supports the accurate prediction of miRNA genes. For the prediction of miRNAs, ncRNA-class Web Tool uses a Support Vector Machine (SVM) classification model proposed in [6] which achieves accuracy over 98%. Furthermore, the tool supports a relational database containing all ncRNA genes data analyzed so far. These data are manually curated by the Web Tool's administrators. Users are able to derive information about their own uploaded non-coding genes or about all non-coding genes maintained in the database through a user friendly search form.

The rest of the paper is organized as follows: Section 2 presents tools for the non coding RNA analysis and for the prediction of miRNA genes focusing on tool with available web interface. Section 3 describes the basic components of our implementation and illustrates the provided services. Section 4 concludes the paper and addresses future directions and enhancements.

2 Existing tools for non-coding RNA feature calculation and for the prediction of miRNA genes

The prediction and the analysis of non-coding RNA genes require the effective calculation of informative sequential, topological and thermodynamical features.

From the late 1980's it was clear that the thermodynamic stability of the RNA molecules is a fingerprint and plays a significant role in their identification. The following years large scale investigations proved that the ncRNAs support a secondary structure which it is significant different from random and other sequences. Many research studies proposed metrics on the RNA secondary structure and they boosted the development of efficient tools such as the UnaFold [7] software, the Vienna RNA package [8] and RNAForester [9]. Despite the fact that many online or standalone tools have been developed for the prediction of informative non-coding RNA features, at present there does not exist any user friendly tool enabling users to calculate most sequential, topological and thermodynamical features at once.

The important role of miRNA genes in the gene regulation mechanism and the limitations of experimental methodologies for the prediction of miRNA genes, have led to the development of a wide variety of computational methods for this purpose [5]. Comparative methods use phylogenetic profiles to exploit the strong conservation of miRNA genes among species. To overcome the bias of phylogenetic profiles in predicting conserved miRNA genes, recent methodologies combine sequential, topological and thermodynamical features deploying machine learning techniques, such as Bayesian classifiers and SVM models.

In contrast to the big number of published computational methods for the prediction of miRNA genes, only a few of them may be accessed through a user friendly interface. The most important of them are presented in Table 1.

MirScan, is a web interface and it is based on simple comparative methods checking whether a candidate gene is similar, in terms of 50 sequential and topological features,

to experimentally verified miRNAs in the organisms of C.elegans and Caenorhabditis briggasae.
MiRAlign, is also a web interface which is based on a comparative method. A similarity score is estimated for every candidate gene using structural and sequential features. The threshold for the prediction of miRNA genes is a user specified variable in order to enable the tuning of the tradeoff between sensitivity and specificity.
MIRcheck, is a public available standalone set of scripts which is specified in the prediction of miRNA genes in plants. This method estimates a similarity score between experimentally verified miRNAs and candidate miRNA genes. This similarity score is based on six features including structural, sequential and conservative features.
miPred may be accessed through a web interface and is based on a machine learning classification algorithm which classifies effectively miRNA genes and pseudo hairpins. The classification algorithm is a SVM model using as inputs sequential, structural and thermodynamical features.

Table 1: Existing online microRNA prediction tools

Name	Url	Reference
MirScan	http://genes.mit.edu/mirscan/	[10]
MiRAlign	http://bioinfo.au.tsinghua.edu.cn/miralign/	[11]
MIRcheck	http://web.wi.mit.edu/bartel/pub/software.html	[12]
miPred	http://www.bioinf.seu.edu.cn/miRNA/	[13]

All existing online or public available tools for the computational prediction of miRNA genes suffer from certain limitations and drawbacks which ncRNAclass Web Tool tries to surpass. Specifically, there does not exist any tool to enable users to check multiple candidate sequences. All of them allow users to insert a single sequence giving as output the result of whether this sequence is a miRNA gene or not. Furthermore, existing tools do not provide information about the calculated feature values. This fact makes the comparison of existing computational methodologies with novel ones a very hard task. Finally, the algorithms which are used in existing online tools do not include any feature selection methodology and the features which are used in each one of them are empirically selected.

3 ncRNAclass Web Tool

NcRNAclass Web Tool integrates a feature calculation module, a miRNA prediction module and a general purpose relational database which should be accessed through a user-friendly interface.

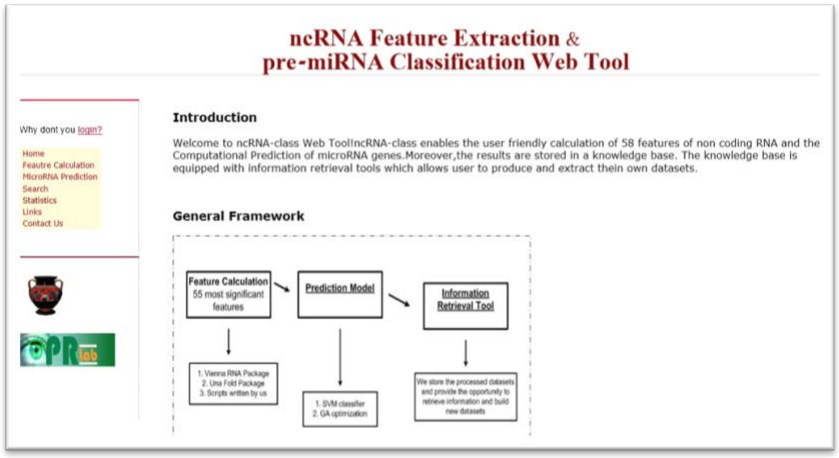

Fig. 1: The main page of ncRNAclass Web Tool

3.1 Feature Calculation Module

The service provided by ncRNA Web Tool is the feature calculation module. The development of algorithms for assessing the statistical significance on attributes contributed in uncovering a number of characteristics capable of predicting novel ncRNAs. In every Machine Learning-based approach, the choice of an appropriate feature set for the training of the classifier is of great importance.

Up to now, tremendous amount of work have been done concerning the choice of discriminate features for the class of the miRNAs. Various feature sets have been proposed, containing information that describes the sequence, the topology and the structure of the hairpins. In our work, we incorporated the features proposed by microPred [14], which is the methodology with the most informative feature set, and by following the literature we introduced some new characteristics that are suitable for the other classes of non-coding RNAs. We have not incorporated features that require phylogenetic filtering, alignment among species or expression profile patterns. The reason is that this kind of information is not always available and it is straightforward to extract this information from the sequences themselves. The final feature set consists of 58 features from various categories. The computation of most of the features is accomplished by using scripts in C language. Regarding the thermodynamical features, functions from the UnaFold and Vienna RNA packages were incorporated. The description of the supported features is the following:

- 16 dinucleotide frequencies (%XY) such that $X,Y \in \Sigma[A,C,G,U]$
- 1 Aggregate Dinucleotide Frequency (%G+C ratio)
- 4 folding measures
    - Adjusted base pairing propensity dP = total_bases/L , where L is the length of the structure and total_bases the number of base pairs in the structure

    - Adjusted Minimum Free Energy of folding dG = MFE/L , where MFE is the minimum free energy of the structure as calculated by the Vienna fold routine
    - Adjusted base pair distance dD
    - Adjusted shannon entropy dQ
- 4 Minimum Free Energy indexes
    - MFE Index 1 = dG/%(C+G)
    - MFE Index 2 = dG/number_of_stems ,where each stem is at least 3 continuous base pairs in the structure
    - MFE Index 3 = dG/number_of_loops , where number_of_loops is the number of the loops in the secondary structure
    - MFE Index 4 = dG/total_bases
- 1 Topological descriptor dF
- 4 RNA fold related features
    - Normalized Ensemble Free Energy
    - Frequency of MFE structure
    - Structural Diversity
    - Diff = |MFE-EFE|/L where, EFE is the ensemble free energy
- 6 UnaFold related features
    - Structure Entropy dS
    - Normalized Structure Entropy dS/L
    - Structure Enthalpy dH
    - Normalized Structure Enthalpy dH/L
    - Melting Temperature Tm
    - Normalized Melting Temperature Tm/L
- 8 Base pair related features
    - |A−U|/L,|G−C|/L, |G−U|/L; where |X −Y| is the number of (X −Y) base pairs in the secondary structure, (X −Y)∈{(A−U), (G−C),(G−U)}.
    - Average base pair per stem
    - %(A−U)/n_stems, %(G−C)/n_stems, %(G−U)/n_stems.
- 4 statistical features : zG,zP,zD,zQ
- 10 new introduced features
    - Length of the sequence
    - Ratio G/C ,where G,C is the number of G,C bases
    - BP/GC,BP/GU,BP/AU, where BP is the total number of base pairs and GC,GU,AU the number of respective base pairs
    - Centroid Energy, Centroid Distance: both of them are RNA folding related attributes and for their calculation the Vienna RNA package is used
    - %(A+U) aggregate frequency
    - MFE Index 5 = dG/%(A+U) ratio
    - Positional Entropy dPs : a new introduced attribute which estimates the structural volatility of the secondary structure

All the aforementioned folding features were computed using the Vienna RNA package functions which calculates the minimum free energy of the RNA structure and provides an estimation of the partition function as described by McCaskill in 1990.

The topological descriptor dF measures the compactness of the secondary structure and it corresponds to the second eigenvalue of the Laplacian matrix which is a

representation of the secondary structure as graph. It was performed using a script written by us according to the RAG publications, instead of using the RAG software.

These thermodynamical features were calculated using routines provided by UnaFold.

In order to study the variance of dG,dP,dD and dQ features we applied the statistical factor z score. Very briefly, the Z-score is the number of standard deviations by which the feature x deviates from the mean of the features x of the set of shuffled sequences. For the purposes of our tool we, use 1000 random sequences generated from the original sequences

According to the biophysical and thermodynamical behavior of the small non coding molecules we added 10 new attributes with high discriminative power. An extensive description of the proposed features can be found in Freyhult et al. and Ding and Lawrence. Also, in these studies the authors measured their discriminative power for the prediction of various ncRNA categories by using simple statistical analysis.

Users are enabled to calculate the full set of 58 features either for their single ncRNA of interest or for a list of ncRNAs written in a Fasta format text document. The calculation starts by typing the number of the sequence or by uploading sequence files in fasta format. The tool using C# functions reads the input of the user and starts the calculations process. This process begins with calculating the thermodynamical features supported by the Unafold software. Then the tool feeds the Vienna RNA package and calculates the rest of the thermodynamical features. The previous results are piped to software written by us that enables the calculations of all the other features and as a final step merges the results. The outcome is piped to our C# form, which displays the results to the screen. Additionally, the tool saves the output to a file and in the background it calls SQL insert functions and stores the results to the database. The next flow chart presents the feature calculation process.

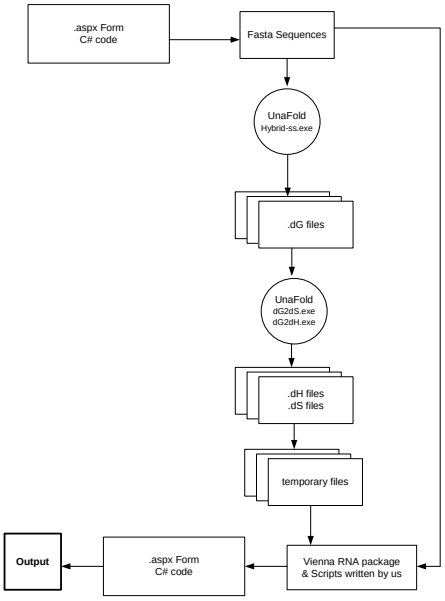

Fig. 2: The feature calculation process

3.2 miRNA Prediction Module: A hybrid methodology

The second module of ncRNAclass Web Tool is the miRNA prediction module. For the microRNA prediction we adopted a novel hybrid approach [6]. This methodology combines the efficiency and robustness of Support Vector Machine classification with Genetic Algorithms in order to classify real human pre-miRNAs from pseudo-hairpins. The SVM classifier was chosen because of its robustness when dealing with noisy and high dimensional data, its good generalization performance [15] and its already proven efficiency in various bioinformatics classification problems. For the effective tuning of the SVM parameters and the feature selection, this hybrid methodology utilizes a Genetic Algorithm.

The experimental results of this methodology led to the extraction of an optimal feature subset. This subset consists of the following features:

- two dinucleotide frequencies (%CG, %GU)
- Minimum Free Energy Index 2 (MFEI2)
- adjusted base pair propensity (dP)
- frequency of MFE structure (Freq)
- Structural diversity (Diversity)
- Structure Entropy to hairpin Length (dS/L)

Experimental results using this feature subset achieved higher performance in terms of sensitivity (0,9928), specificity (0,9927)and accuracy (0.9927). Using this model we implemented a miRNA predictor and we incorporated the executable program in our web tool platform.

The software was programmed using the Matlab R2009b and it is able to handle files containing more than one sequence in fasta format. The communication between the prediction software and the web tool was performed using routines written in C# language. The prediction results can be viewed online or can be downloaded as txt files.

3.3 ncRNAclass Web Tool's Database

The last component of the ncRNAclass Web Tool is a relational database for storing all the processed sequences. A very simple design schema was used containing all ncRNA genes data analyzed so far. These data are manually curated by the Web Tool's administrators. Users are given the opportunity to derive information about their own uploaded non-coding genes or about all non-coding genes maintained in the database through a user friendly search form.

The database at present supports the storage of miRNAs in three different tables: the predicted miRNAs, pseudo hairpins and unspecified non-coding RNAs. Every tuple in the database is keyed by a unique id and for every calculated feature a table attribute of float data type is stored. Statistical information is also maintained that can be used for more complex aggregate operations. The simple design of our database enables the retrieval of significant information and our tool provides an easy way of accessing. The users have many abilities to ask queries in the databases. First and

foremost, they can search the supported tables by using the name of the organisms they are interested in. More advances search option provide search by sequence or by specifying the time period. The latter provided service is very beneficial and the user can keep track of the updates in the database and download the latest processed sequences. Also it gives us the opportunity to retrieve complex statistical information. The registered users in the site have additional operations. They are able to upload files and after the processing step they can manage the information they are interested in. In simple words the database acts like a local warehouse for the users of the system. Moreover the tool provides multiple search options by combining all the above criteria. The implementation of the database performed in SQL and the interface programmed using C# under the Asp.net environment.

4 Discussion

The ncRNA-class Web Tool aims to promote the research about non-coding genes. It provides the capability to calculate several important ncRNA features predicts miRNA genes and enables information mining about non-coding genes through its database. In opposite to existing methodologies, it enables the analysis of many sequences in a single run and calculates at once the most informative features which have been proposed in the literature. Furthermore, for the prediction of miRNAs, it deploys a modern method which presents extremely high classification results.

Our future plans involve the incorporation of prediction tools for other ncRNA types into the ncRNA-class Web Tool. To satisfy this goal, we plan to incorporate some features proposed in the literature which are specific for some categories of non-coding RNAs such as t-RNA and sno-RNA. At present we are incorporating a new module in ncRNAclass Web Tool, which will enable users to find target genes for every miRNA gene using the methodology proposed in [16]. This module will also support the calculation of the features which were used for the prediction of the target genes. Finally, our future research agenda includes the optimization of the execution pipeline in order to reduce the running time and achieve greater performance for huge amounts of data.

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. Mattick, J. S.: Non-coding RNA. Human Molecular Genetics. 15(90001), R17--R29, (2006)
2. Esteller, M.: Non-coding RNAs in human disease. Nat Rev Genet. 12 (12), 861--874 (2011)
3. Bartel, D. P.: MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 116(2), 281--297, (2004)

4. Lai, E.C.: microRNAs: runts of the genome assert themselves. Curr. Biol. 13(23), R925—936 (2003)
5. Mendes, N.D., Freitas, A.T. and Sagot M.F.:Current tools for the identification of miRNA genes and their targets. Nucleic Acids Research. 37(8), 2419 --2433, (2009)
6. Theofilatos, K., Kleftogiannis, D., Rapsomaniki, M., Haidinis, V., Likothanassis, S., Tsakalidis, A., Mavroudi, S. A novel embedded pre-miRNA classification approach for the prediction of microRNA genes. In proceeding of the 10th IEEE International conference on Information Technology and Application in Biomedicine (ITAB 2010), Corfu, Greece (2010)
7. Markham, N.R., Zuker, M.: UNAFold: software for nucleic acid folding and hybridization. Methods Mol. Biol. 453, 3--31 (2008)
8. Hofacker, I. et al.: Vienna RNA secondary structure server. Nucleic Acids Research. 31(13), 3429--3431 (2003)
9. Höchsmann, M., Voss, B. and Giegerich, R.: Pure Multiple RNA Secondary Structure Alignments: A Progressive Profile Approach. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 1(1), 53—62 (2004)
10. Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., and Bartel, D.P.: The microRNAs of Caenorhabditis Elegans. Genes Dev. 17(8), 991—1008 (2003)
11. Wang, X., Zhang, J., Li, F., Gu, J., He, T., Zhang, X., and Li, Y.: MicroRNA Identification Based on Sequence and Structure Alignment. Bioinformatics. 21(18), 3610--3614 (2005)
12. Jones-Rhoades, M.W. and Bartel, D.P.: Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Mol. Cell. 14, 787--799 (2004)
13. Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X. and Lu, Z.: MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res. 35, W339--W344 (2007)
14. Batuwita, R. and Palade V.: microPred: effective classification of pre-miRNAs for human miRNA gene prediction. Bioinformatics. 25, 989-995 (2009)
15. Burges, C.J.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery. 2, 121--167 (1998)
16. Korfiati, A., Kleftogiannis, D., Theofilatos, K., Likothanassis, S., Tsakalidis, A. and Mavroudi, S. Predicting Human miRNA Target Genes Using a Novel Evolutionary Methodology. In proceeding of the 7th Hellenic Conference on Artificial Intelligence (SETN2012) Lamia, Greece, (2012)