A PROBABILISTIC METHOD FOR PREDICTION OF MICRORNA-TARGET INTERACTIONS

Hasan Oğul[1], Sinan U. Umu[2,3], Y. Yener Tuncel[3] and Mahinur S. Akkaya[2]
[1]*Department of Computer Engineering, Başkent University, Ankara, Turkey*
[2]*Department of Chemistry, Middle East Technical University, Ankara, Turkey*
[3]*Bioinformatics Program, Informatics Institute, Middle East Technical University, Ankara, Turkey*

Keywords: MicroRNA target prediction, Markov chains.

Abstract: Elucidation of microRNA activity is a crucial step in understanding gene regulation. One key problem in this effort is how to model the pairwise interaction of microRNAs with their targets. As this interaction is strongly mediated by their sequences, it is desired to set up a probabilistic model to explain the binding between a microRNA sequence and the sequence of a putative target. To this end, we introduce a new model of microRNA-target binding, which transforms an aligned duplex to a new sequence and defines the likelihood of this sequence using a Variable Length Markov Chain. It offers a complementary representation of microRNA-mRNA pairs for microRNA target prediction tools or other probabilistic frameworks of integrative gene regulation analysis. The performance of present model is evaluated by its ability to predict microRNA-mRNA interaction given a mature microRNA sequence and a putative mRNA binding site. In regard to classification accuracy, it outperforms a recent method based on support vector machines.

1 INTRODUCTION

MicroRNAs (MiRNAs) are tiny regulators of gene expression in post-transcriptional level. Many evidences suggest that they abundantly exist in many organisms and play pivotal roles in regulation of several biological processes (Lee et al., 1993); (Bartel, 2004). Last decade of bioinformatics research has been strongly influenced by the discovery of miRNAs. Recent findings about the complex behaviour of miRNAs have triggered the application of computational methods to analyse miRNA functions and activities. Relevant computational research has been enriched in several directions like predicting miRNA targets (Alexiou et al., 2009); (Saito et al., 2010); (Mendes et al., 2009); (Hammell 2010); (Bartel, 2009), inferring regulatory modules or networks comprising miRNAs (Yoon and Micheli, 2005; Liu et al. 2010), identifying disease-related miRNAs (Lu et al., 2005); (Madden et al., 2010) or finding functionally related miRNAs (Yu and He, 2011); (Wang et al., 2010). The studies have been resulted with various online tools or standalone programs developed for storage, retrieval or analysis of miRNA related data.

Despite the enormous number miRNA target prediction programs, current systems biology can benefit too little from their result (Barbato et al., 2009). The reasons for the limitation of available tools are two-fold. First, it is hard to see a consensus in their prediction results, which make them unreliable to use in generation of further hypotheses. Second, a conventional tool for target prediction can only reveal a one-to-one relationship between a miRNA and an mRNA under consideration. However, we know that the picture is bigger: one miRNA can regulate many genes and one gene can be regulated by multiple miRNAs in cooperation, even with other regulatory factors (Krek et al., 2005; Lim et al., 2005). An obvious solution to these problems is to integrate multiple data sources to infer more reliable and native explanations. The lack of sufficient data samples will also motivate the use of probabilistic methods to construct integrative frameworks to analyse the multi-source data. Since it is well-known that a miRNA can target an mRNA in a sequence-specific manner (Bartel, 2004), an integrative model, whatever its practical purpose is, should incorporate the sequence information. Furthermore, such a model of sequence-based miRNA-mRNA binding should be defined in a

Oğul H., U. Umu S., Yener Tuncel Y. and S. Akkaya M..
A PROBABILISTIC METHOD FOR PREDICTION OF MICRORNA-TARGET INTERACTIONS.
DOI: 10.5220/0003694902890293
In *Proceedings of the International Conference on Neural Computation Theory and Applications* (NCTA-2011), pages 289-293
ISBN: 978-989-8425-84-3

probabilistic way so that it could represent the degree of this interaction and be easily incorporated into other data. Recent attempts for integrating sequence with functional data have largely ignored this issue and considered the sequence-directed miRNA-mRNA interaction in a binary way, which is usually obtained from other target predictors.

In this study, we aim to provide a means of modelling miRNA and mRNA regulatory relationship by a probabilistic description over the sequential content of a putative miRNA-mRNA duplex. In this respect, our model first performs a complementary alignment between the mature miRNA sequence and a putative binding site from present mRNA. Resulting alignment is represented by a new sequence over an alphabet of possible matches or mismatches, where different base paring rules are taken into account by distinct alphabet symbols. The probability of new sequence of miRNA-mRNA duplex is then analysed using a Variable Length Markov Chain (VLMC) approach. VLMC (Ron et al., 1996) is a flexible yet powerful model to analyse a sequential content based on the order of local arrangements by quantifying the probability of the occurrence of a specific symbol after a certain sub-sequence with varying length less than a predefined maximum. This enables us to calculate the likelihood of whole sequence by simply multiplying local probabilities. For miRNA-mRNA duplex case, the independence of model from global position information allows the evaluation of significant local regions which might be enriched in any part of miRNA-target duplex formation. Therefore, the order of distinct base pairs and mismatches are taken into account in addition to their frequency of appearances. Two VLMC-based likelihoods of new sequence which are obtained from positive and negative training sets can reveal the degree of a potential interaction between two entities.

```
5' GGGUGUUAAGACUUGACACAGUACCUCG 3'
   .. |.| | .||.| ||    |||||||.
3' UUGAUA-UGUUGGA-UG----AUGGAGU 5'

   eegbebqageadfbqadqqqqbaccbce
```

Figure 1: Alignment of miRNA sequence (middle) and mRNA binding site (top) is transformed into a new sequence of duplex formation (bottom). It is defined over a new alphabet of eight letters; six for representing all possible directional base pairings including G-U and U-G wobbles, one for mismatches and one for spaces in any other site.

2 METHODS

2.1 miRNA-mRNA Duplex Sequence Model

MiRNA-mRNA duplex is constructed by complementary alignment of mature miRNA sequence with mRNA binding site. Resulting alignment transformed into a new sequence defined over an alphabet of symbols representing a distinct nucleotide pair type including mismatches and spaces in any other site (Figure 1). The alignment algorithm employs well-known dynamic programming algorithm of Smith and Waterman (1981).

2.2 Markov Chain Model

Markov chains are used to model sequential data in terms of order of individual symbols but regardless of their global positions. A simple model to define the likelihood of a sequence S_1^N is to assume a zero-order Markov Chain and compute the probability by multiplying the probabilities of each symbols contained, i.e.,

$$P(S_1^N) = \prod_{j=1}^{N} P(S_j = s_j)$$

where P(.) refers to probability, S_j is the random variable representing the letter at position j with s_j as its realization.

In Markov chain model, the probability is identical for each position, thus ignoring the order of amino acids. The model assumes that each position is independent from the others; the context information is not taken into account at all. Since the complex structure of biological sequences is poorly reflected with this approach, the use of higher order models has been suggested (Thijs et al., 2001). In this respect, the likelihood of the sequence can be defined by an L-order Markov chain, where the probability of each symbol depends on the preceding subsequence of a fixed length $L<N$.

Fixed-order Markov models, although able to modeling rich sources in an efficient manner, have critical drawbacks for practical use. The major problem is that the number of model parameters grows exponentially, resulting in a very sharp and discontinuous transition from under-fitted models to over-fitted models. Therefore, for a trivial model order, the model suffers from learnability hardness results, and consequently, the derived model is not guaranteed to be optimal (Bejerano and Yona, 2001;

Ben-Gal et al., 2005). Another limitation is that a fixed length model order ignores the domain specific nature of biological sequences and, even with an optimized selection of L value, the model may not be effective in the detection of significant cut-off locations in the duplex chain. A more flexible version of higher order Markov models allows a variable order, i.e. memory length, that depends on the preceding subsequence to given position such that the order of the model becomes a function the context at each position (Ron et al., 1996). We further extend the model to take into account the succeeding subsequence and define the sequence likelihood as

$$P(S_1^N) = \prod_{j=1}^{N} P(S_j = s_j \mid S_{j-L_j}^{j-1} = s_{j-L_j}^{j-1}).P(S_j = s_j \mid S_{j+L'_j}^{j+1} = s_{j+L'_j}^{j+1})$$

where L_j and L'_j are the optimal lengths for preceding and succeeding subsequences respectively. The last modification enables considering the context surrounding the symbol and provides a better generalization of the model.

2.3 miRNA-mRNA Interaction Prediction

MiRNA-mRNA interaction prediction can be considered to be a binary classification problem where a duplex sequence is required to be assigned to one of the positive ($C=1$) or negative ($C\neq1$) classes. Given a sequence $S_1^N = s_1 s_2 ... s_N$, starting from the position 1 and ending at the position N, where s_i is drawn from the finite alphabet of nucleotide pair symbols, we deploy a classification rule based on the likelihood of the S_1^N that reports a positive (target interaction) label if $P(S_1^N \mid C = 1) > P(S_1^N \mid C \neq 1)$ a negative (no interaction) label otherwise. The conditional probabilities for positive and negative cases are inferred from corresponding training sets.

3 RESULT

MiRNA target prediction methods have often used validated data together with some predicted data either to develop or assess their algorithms in order not to suffer from small sample size. When an evaluation of target prediction ability for given miRNAs and genes is in question, it is reasonable to ignore the lack of verified binding site information since they only deal with the binary result. However, a correct and reliable evaluation of a probabilistic binding model requires a set of experimentally verified binding sites, both for positive and negative cases. In a recent study of Yang et al. (2008), they collected a number of experimentally validated positive and negative miRNA-mRNA duplex sets. The repository contains 233 miRNA-binding site pairs from drosophila, C. elegans, human, mouse, rat and zebrafish, where 195 of them are positive interactions and 38 are negatives. Negative samples still contain at least six perfect base pairs in seed region for reliable evaluation of the algorithms. Note that a positive interaction is referred as targeting relationship between miRNA and mRNA under consideration in this context. In the same publication they propose an algorithm, called miRTif, to predict the positive or negative interaction of given duplex based on a powerful machine learning technique called Support Vector Machines. They reported the sensitivity (percentage of actual positives correctly predicted as such), specificity (percentage of actual negatives correctly predicted as such), accuracy (percentage of all correctly predicted samples) and AUC (Area under Receiver Operating Characteristics Curve) of their predictions based on 10-fold cross validation tests. To make a fair comparison, we compiled the experiments in a 10-fold cross validation set-up on the same data set. Since the selection of arbitrary 10 groups might produce different results, we run the algorithm several times for different initial configurations and took their average. Results of the experiments are shown in Table 1 in comparison with miRTif.

Table 1: Comparison of present model with miRTif on miRNA-mRNA interaction prediction.

Method	miRTif	present method
Sensitivity	83.6%	86.7%
Specificity	73.7%	73.7%
Accuracy	81.9%	84.6%
AUC	0.89	0.94

During this experiment, we treated the miRNA-mRNA duplex as a sequence over a 8-letter alphabet, where each directed base pair, including wobbles, i.e. A-U, U-A, G-C, C-G, G-U and U-G, was represented by a distinct letter and all mismatches were unified into one letter. An additional letter was used for a space in any other site. Maximum length for Markov chains was set to 5 since longer motifs are quite rare, which may not contribute to the result, but several significant motifs could be observed up to 5-letter length. We ignored

the single appearance of any sub-string in positive or negative set to eliminate the data set bias.

The table demonstrates that new method can outperform miRTif in all evaluation criteria except specificity, where they perform equally. The superiority of present method is quite significant in terms of ROC statistics. This shows that new method is not only successful in separation of positive and negative examples but also able to successfully quantize the level of certainty in its prediction. This result is consistent with our intention to release this model; enabling the sequence to be easily integrated with other data.

4 CONCLUSIONS

We have introduced a probabilistic method to model miRNA-target binding and evaluated its performance on prediction of the interaction when miRNA sequence and a putative binding site are given. The accurate results that we obtained from the experiments suggest that present model is able to capture compositional properties of a duplex sequence by additionally considering the effect of different base pairings, mismatches and gaps with their arrangements inside the duplex.

The model which we proposed may find applications in several platforms. First, it can be used as a post-processing filter for other miRNA target prediction tools. Many of available algorithms consider the seed match as a strong evidence for target identification. Since it is possible to observe random mRNA matches to seed region without any interaction, this decision criterion can mislead the algorithm to produce excessive number of false positives. Present method is able to reject miRNA-nontarget duplexes despite a high seed complementarity, thus it may help to reduce the number of false positives by reanalysing the binding site predicted by former tool. Second, it may serve complementary information which can be deployed in target prediction algorithms. Conventional methods perform a window-based linear scan over the mRNA sequence to identify a putative binding site which may attain a large binding score based on a weighted sum of predefined criteria. Output of proposed model is an obvious complement to other determinants such as structure, site accessibility or cross-spices conservation in this scoring scheme. Third, the model enables the researchers to integrate sequence data directly with other behavioural data such as gene expression profiles over a probabilistic framework. An integrated framework can provide a comprehensive analysis of miRNA functions associated with other entities, conditions or diseases. Machine learning research has been competing in two directions for intelligent analysis of heterogeneous data: black-box kernel methods such as Support Vector Machines and probabilistic graphical models such as Bayesian Networks. Latter requires a probabilistic representation of each contributor in the model. Present scheme can fill a gap in this respect.

It is anticipated that the participation of computational models into miRNA research will increasingly continue in coming years. We believe that integration of multi-source heterogeneous data will be a focal point in this research. Our study does not yield a standalone tool in this context; however, it provides a different view of miRNA-target interactions from which future research can definitely benefit. As a future work, we plan to analyze the effects of seed and non-seed regions and the types of different pairings of match and mismatches in duplex analysis. Our final goal is to come up with an integrative solution which combines this sequence-based model with other behavioural data in order to find functional maps of miRNAs and their targets.

ACKNOWLEDGEMENTS

This study was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under the Project 110E160.

REFERENCES

Alexiou P., Maragkakis M., Papadopoulos G. L. et al., Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics* 2009; 25:3049-3055.

Barbato C, Arisi I, Frizzo M. E. et al. Computational Challenges in miRNATarget Predictions: To be or Not to be a True Target? *J Biomed Biotechnol* 2009; 2009:803069.

Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell* 2009; 136: 215-233.

Bartel D. P., MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004;116:281-297

Begleiter R., El-Yaniv R., Yona G., On prediction using variable order Markov models, Journal of Artificial Intelligence Research 2004, 22:385-421.

Bejerano G., Yona G., 2001. Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. *Bioinformatics* 17, 23-43.

Ben-Gal I., Shani A., Gohr A., Grau J., Arviv S., Shmilovici A., Identification of transcription factor binding sites with variable-order Bayesian networks, *Bioinformatics* 2004; 21: 2657-2666.

Buhlman P., Wyner A. J., Variable length Markov chains, *Annals of Statistics* 1999; 27: 480-513.

Hammell M., Computational methods to identify miRNA targets. *Semin Cell Dev Biol* 2010 Sep;21(7):738-44.

Krek A., Grun D., Poy M. N., Wolf R., Rosenberg L., Epstein E. J., MacMenamin P., da Piedade I., Gunsalus K. C., Stoffel M., Rajewsky N., Combinatorial microRNA target predictions. *Nat.Genet.* 2005;37:495-500.

Lee R. C., Feinbaum R. L., Ambros V., The C. Elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 1993; 75: 843-854.

Liu B., Liu L., Tsykin A et al. Identifying functional miRNA–mRNA regulatory modules with correspondence latent dirichlet allocation. *Bioinformatics* 2010; 26:3105-3111.

Lu J, Getz G, Miska EA et al. MicroRNA expression profiles classify human cancers. *Nature* 2005; 435: 834–838.

Madden SF, Carpenter SB, Jeffery IB et al. Detecting microRNA activity from gene expression data, *BMC Bioinformatics* 2010; 11:257.

Mendes ND, Freitas AT, Sagot MF. Current tools for the identification of miRNA genes and their targets. *Nucleic Acid Res* 2009; 37: 2419-2433.

Ron D., Singer Y., Tishby N., The power of amnesia: Learning probabilistic automata with variable memory length, *Machine Learning* 1996; 25: 117-149.

Saito T, Saetrom P. MicroRNAs - targeting and target prediction. New Biotechnology 2010; 27: 243-249.

Smith T and Waterman M, "Identification of common molecular subsequences", *Journal of Molecular Biology* 1981, 147: 195-7.

Thijs G., Lescot M., Marchal K., Ronbauts S., De Moor B., Rouze P., Moreau Y., A higher order background model improves the detection of promoter regulatory elements by Gibbs sampling, *Bioinformatics* 2001; 17: 1113-1122.

Wang D., Wang J., Lu M., Song F. and Cui Q., Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases, *Bioinformatics* 2010, 26(13): 1644-1650.

Yang Y., Wang Y. P., Li, K. B., MiRTif: a support vector machine-based microRNA target interaction filter. *BMC Bioinformatics* 2008; 9:S4.

Yoon S. and Micheli G., Prediction of regulatory modules comprising microRNAs and target genes, *Bioinformatics* 2005; 21: i93-i100.

Yu G. and He Q., Functional similarity analysis of human virus-encoded miRNAs, *Journal of Clinical Bioinformatics* 2011; 1:15.