Predicting miRNA targets utilizing an extended profile HMM

Jan Grau[1,*], Daniel Arend[1], Ivo Grosse[1], Artemis G. Hatzigeorgiou[2], Jens Keilwagen[3], Manolis Maragkakis[1,2], Claus Weinholdt[1], and Stefan Posch[1]

[1] Institute of Computer Science, Martin Luther University Halle–Wittenberg, Germany
[2] Institute of Molecular Oncology, Biomedical Sciences Research Center 'Alexander Fleming', Vari, Greece
[3] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

Abstract: The regulation of many cellular processes is influenced by miRNAs, and bioinformatics approaches for predicting miRNA targets evolve rapidly. Here, we propose conditional profile HMMs that learn rules of miRNA-target site interaction automatically from data. We demonstrate that conditional profile HMMs detect the rules implemented into existing approaches from their predictions. And we show that a simple UTR model utilizing conditional profile HMMs predicts target genes of miRNAs with a precision that is competitive compared to leading approaches, although it does not exploit cross-species conservation.

1 Introduction

miRNAs are short ($\sim$ 22 nt) endogeneous RNAs that bind to partially complementary sites on mRNA target sequences. They induce cleavage of the miRNA-mRNA duplex or repress translation of the bound mRNA [BSRC05]. Hence, miRNAs influence gene expression and introduce a novel level of gene regulation. For instance, several miRNA signatures have already been successfully associated with human cancers. In animals, miRNAs preferentially bind to the 3' untranslated region (UTR) of the mRNA, and for binding a high complementarity between miRNA and target is required only at the 5' end of the miRNA. Computational miRNA target prediction plays a key role in deciphering the functional role of miRNAs. Several dozen programs have been therefore developed in the last years, and in the following, we describe the main idea behind some of the most widely used programs.

[LSJR$^+$03] propose an algorithm for the prediction of targets of vertebrate miRNAs called TargetScan. TargetScan requires perfect complementarity between positions 2 and 8 at the 5'-end of the miRNA and a potential target, and the free energy of binding between miRNA and target is computed. Predictions are verified using orthologous UTR sequences from other organisms. [LBB05] propose a refined version called TargetScanS, which demands a shorter region of the target to be complementary to nucleotides $2-7$ of the miRNA. TargetScan 5.0 [FFBB09] additionally considers the distance from the 3' UTR and AU content.

In contrast to TargetScan, miRanda [EJG+03] does not require perfect complementarity at the seed region, but uses an algorithm similar to Smith-Waterman sequence alignment with similarity scores of $+5$ for G:C and A:U basepairs, $+2$ for G:U basepairs, and -3 for mismatches, and the scores for the first 11 positions of the alignment are weighted by a factor of 2. Potential target sites (TSs) are filtered for a minimum similarity score and a minimum free energy.

PicTar [KGP+05] searches for perfectly complementary seed regions of 7 nt starting from position 1 or 2 of the miRNA. Mismatches in the seed region are allowed if these do not increase the free energy. Additionally, a filter with respect to the free energy of the complete miRNA-mRNA duplex is applied.

DIANA-microT [MRS+09] prefers perfect complementarity of 7 to 9 nt starting from position 1 or 2 of the miRNA. However, if the considered TS shows good complementarity to the 3' end of the miRNA, the length of this seed region may be reduced to 6 nt, and single G:U basepairs are allowed. DIANA-microT uses orthologous UTRs from up to 27 organisms for assessing the conservation of TSs. Finally, the score of a potential UTR target is computed as a weighted average of all predicted TSs.

In contrast to previous approaches, we propose a fully statistical approach for predicting TSs of given miRNAs that is capable of learning rules of miRNA-TS binding from data sets comprising pairs of miRNAs and associated TSs. This approach employs an extension of profile hidden Markov models (HMMs) [KBM+94], which we call *conditional profile HMM* (CoProHMM), and learns parameters by the discriminative maximum supervised posterior (MSP) principle [CdM05, GKK+07]. Since all parameters of CoProHMMs are learned from training data, this approach is not biased towards heuristic assumptions about miRNA-TS interaction like the existence or length of a seed region.

2 Methods

In the following, we introduce CoProHMMs for modeling the binding between miRNA and TS. We describe how we learn CoProHMMs from data, and explain how we combine several predictions of a learned CoProHMM to predict target genes of a given miRNA.

2.1 Conditional profile HMMs

At the basis of the CoProHMM modeling miRNA TSs, we use a standard profile HMM architecture [KBM+94], which is illustrated in Fig. 1. This architecture is also referred to as "plan9" due to its 9 transitions at each layer of the model. We define a total of K match states M_k, which emit a nucleotide of the TS with a probability that is conditional on the nucleotide at position k of the miRNA. Here, we use $K = 22$, since this is the length of a typical miRNA and, hence, the model covers all positions of the miRNA that are potentially interacting with the TS. If a TS and the associated miRNA are perfectly complementary, we anticipate that only match states are visited for emitting the complete sequence of the TS. Otherwise, silent delete states D_k allow for the insertion of gaps into

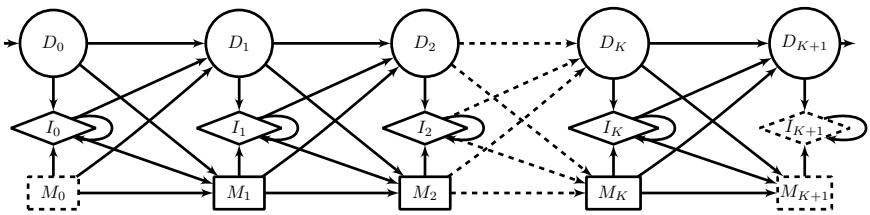


Figure 1: Plan9 architecture of the proposed CoProHMMs. Circles represent silent delete states that do not emit nucleotides of the TS, diamonds represent insert states that emit nucleotides of the TS without considering the nucleotides of the miRNA, and rectangles represent match states that emit nucleotides of the TS with probabilities conditional on the nucleotides of the miRNA. Admissible paths start at D_0 and end at D_{K+1}. States with dashed borders are not visited in admissible paths.

the TS, insert states I_k allow for including gaps in the miRNA, and match states also allow to replace nucleotides. In Fig. 1, edges represent transition probabilities not fixed to 0. From each node of column k, we can reach node I_k in the same column, and nodes M_{k+1} and D_{k+1} in the next column. Each admissible path starts at D_0 and ends at D_{K+1}. Hence, the states M_0, I_{K+1}, and M_{K+1} are never visited in admissible paths, and are only included to simplify recursive definitions in the following.

We parameterize the transition probabilities and the emission probabilities by normalized exponentials [Mac98, BB01] using real-valued parameters, since this allows for an unconstrained numerical optimization of the parameters with respect to the discriminative MSP principle.

According to the plan9 architecture, we define the transition probability $P_T(V|S_k, \boldsymbol{\beta}_{T,S_k})$ of going from node $S_k \in \{I_k, M_k, D_k\}$ to node V given parameters $\boldsymbol{\beta}_{T,S_k}$ as

$$P_T(V|S_k, \boldsymbol{\beta}_{T,S_k}) = \begin{cases} \frac{\exp(\beta_{V|S_k})}{\sum_{\tilde{V}\in\{I_k, M_{k+1}, D_{k+1}\}} \exp(\beta_{\tilde{V}|S_k})} & \text{if } V \in \{I_k, M_{k+1}, D_{k+1}\} \\ 0 & \text{otherwise} \end{cases},$$

where $\boldsymbol{\beta}_{T,S_k} = (\beta_{I_k|S_k}, \beta_{M_{k+1}|S_k}, \beta_{D_{k+1}|S_k}), \beta_{V|S_k} \in \mathbb{R}$.

In contrast to standard profile HMMs, we use conditional probabilities depending on the nucleotides of the miRNA for the emissions of the match states. For match state M_k, we define the conditional emission probability $P_{M_k}(a|r_k, \boldsymbol{\beta}_{M_k})$ of symbol a in the TS given the k-th symbol r_k of the miRNA and parameters $\boldsymbol{\beta}_{M_k}$ as

$$P_{M_k}(a|r_k, \boldsymbol{\beta}_{M_k}) = \frac{\exp(\beta_{a|r_k,M_k})}{\sum_{\tilde{a}\in\Sigma} \exp(\beta_{\tilde{a}|r_k,M_k})}, \tag{1}$$

where $\boldsymbol{\beta}_{M_k} = (\beta_{A|A,M_k}, \beta_{C|A,M_k}, \ldots, \beta_{U|U,M_k}), \beta_{a|b,M_k} \in \mathbb{R}$. Finally, we parameterize the emission probability $P_{I_k}(a|\boldsymbol{\beta}_{I_k})$ of symbol a at insert state I_k given parameters $\boldsymbol{\beta}_{I_k}$ in analogy to equation (1).

We define *forward* variables $\mathcal{F}_{S_k}(\ell, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta})$ as the probability of observing the first ℓ symbols of the TS sequence $\boldsymbol{x}$ and visiting node S_k in state interval $s(\ell, \boldsymbol{x}|\boldsymbol{r})$ given parameters $\boldsymbol{\beta}$ and the sequence $\boldsymbol{r}$ of the miRNA, i.e.,

$$\mathcal{F}_{S_k}(\ell, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta}) = P\left(x_1, \ldots, x_\ell, S_k \in s(\ell, \boldsymbol{x}|\boldsymbol{r})|\boldsymbol{r}, \boldsymbol{\beta}\right). \tag{2}$$

A node S_k is visited in state interval $s(\ell, \boldsymbol{x}|\boldsymbol{r})$ if it is contained in a path from D_0 to D_{K+1}, and the symbols x_1 to x_ℓ have been emitted either by predecessors of S_k in the path or by S_k itself, whereas $x_{\ell+1}$ is emitted by a successor of S_k in this path. We use these forward variables for defining the likelihood $P(\boldsymbol{x}|ts, \boldsymbol{r}, \boldsymbol{\beta}_{ts})$ of TS $\boldsymbol{x}$ given the class ts of TS, the sequence of the miRNA $\boldsymbol{r}$, and parameters $\boldsymbol{\beta}_{ts}$, i.e.

$$P(\boldsymbol{x}|ts, \boldsymbol{r}, \boldsymbol{\beta}_{ts}) = \mathcal{F}_{D_{K+1}}(L, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta}_{ts}). \tag{3}$$

Using this definition, the likelihood $P(\boldsymbol{x}|ts, \boldsymbol{r}, \boldsymbol{\beta}_{ts})$ is not necessarily normalized over all possible sequences $\boldsymbol{x} \in \Sigma^L$ of given length L.

Similar to original profile HMMs, we recursively derive the forward variables of match state M_k using its predecessors $S_{k-1} \in \{I_{k-1}, D_{k-1}, M_{k-1}\}$ from the previous column of the plan9 architecture (cf. Fig. 1) as

$$\begin{aligned}\mathcal{F}_{M_k}(\ell, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta}) = P_{M_k}(x_\ell|r_k, \boldsymbol{\beta}_{M_k}) \\ \sum_{S_{k-1}} \mathcal{F}_{S_{k-1}}(\ell-1, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta})\ P_T(M_k|S_{k-1}, \boldsymbol{\beta}_{T,S_{k-1}}).\end{aligned} \tag{4}$$

In analogy, we derive the forward variables of insert states and delete states.

We initialize the forward variables as follows: We can observe D_0 only before the emission of the first symbol. Hence, we set $\mathcal{F}_{D_0}(\ell, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta})$ to 1 if $\ell = 0$ and to 0 otherwise. We cannot reach M_0 in any admissible path and, thus, $\mathcal{F}_{M_0}(\ell, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta}) = 0$. Finally, we set $\mathcal{F}_{S_k}(0, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta}) = 0$ for all emitting states S_k.

2.2 Discriminative training

For learning the parameters of the CoProHMM discriminatively, we need an additional background model. Here, we use a homogeneous Markov model of order 1 with parameters $\boldsymbol{\beta}_{bg}$ that do not depend on the miRNA $\boldsymbol{r}$, i.e.,

$$P(\boldsymbol{x}|bg, \boldsymbol{r}, \boldsymbol{\beta}_{bg}) = P_{hMM(1)}(\boldsymbol{x}|\boldsymbol{\beta}_{bg}). \tag{5}$$

We derive the class posterior of class $c \in \{ts, bg\}$ using the likelihoods $P(\boldsymbol{x}|c, \boldsymbol{r}, \boldsymbol{\beta}_c)$ of equations (3) and (5) as

$$P\left(c \,|\, \boldsymbol{x}, \boldsymbol{r}, \boldsymbol{\beta}\right) = \frac{P(c|\boldsymbol{\beta})P(\boldsymbol{x}|c, \boldsymbol{r}, \boldsymbol{\beta}_c)}{\sum_{\tilde{c}} P(\tilde{c}|\boldsymbol{\beta})P(\boldsymbol{x}|\tilde{c}, \boldsymbol{r}, \boldsymbol{\beta}_{\tilde{c}})}, \tag{6}$$

where $P(c|\boldsymbol{\beta})$ denotes the a-priori probability of class c, which we parameterize in analogy to equation (1).

For Bayesian inference, we define a prior on the parameters $\boldsymbol{\beta}$. For the homogeneous Markov model of class bg, we use a transformed product-Dirichlet prior [Mac98] with equivalent sample size (ESS) [HGC95] $\alpha_{bg} \cdot K$. We define another transformed product-Dirichlet prior with ESS α_{ts} for the parameters of the CoProHMM, which is the product of independent transformed Dirichlet priors for each set of transition parameters and each set of emission parameters. We use Dirichlet priors, since these are conjugate to the likelihood of the homogeneous Markov model and to the distribution of transitions and (conditional)

emissions. Hence, their hyper-parameters can be intuitively interpreted as pseudo counts. In the following studies, we use $\alpha_{bg} = \alpha_{ts} = 4$.

We learn all parameters $\boldsymbol{\beta}$ on a set of labelled training data $(\boldsymbol{x}_1, \boldsymbol{r}_1, c_1), \ldots, (\boldsymbol{x}_N, \boldsymbol{r}_n, c_N)$. These training data comprise a sufficient number of TSs, i.e. $c_n = ts$, and non-TSs of several miRNAs. Learning the parameters on the TSs of multiple miRNAs conjointly is motivated by the expectation that by this means, CoProHMM may detect general rules of miRNA-TS binding, that could not be detected if we, for instance, learned a standard profile HMM on the TSs of a single miRNA.

We optimize the parameters with respect to the discriminative MSP principle [CdM05, GKK+07], i.e.,

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left[\prod_{n=1}^{N} P\left(c_n \mid \boldsymbol{x}_n, \boldsymbol{r}_n, \boldsymbol{\beta}\right) \right] q\left(\boldsymbol{\beta} \mid \alpha_{bg}, \alpha_{ts}\right), \tag{7}$$

where $q\left(\boldsymbol{\beta} \mid \alpha_{bg}, \alpha_{ts}\right)$ denotes the product-Dirichlet priors on the parameters $\boldsymbol{\beta}$. This optimization must be carried out numerically, which we accomplish by a quasi-Newton second order method.

2.3 Predicting target genes

In the following, we describe how we utilize a CoProHMM for predicting target genes of a miRNA $\boldsymbol{r}$. We assume that the CoProHMM has already been trained on a set of miRNAs – not necessarily including $\boldsymbol{r}$ – and associated TSs and non-TSs. To this end, we extract the UTR $\boldsymbol{y}_n$ of each gene n. Using a sliding window of width $|\boldsymbol{r}|$, we apply the CoProHMM to each sub-sequence of $\boldsymbol{y}_n$ and compute the log-likelihood according to equation (3) given miRNA $\boldsymbol{r}$. For each UTR, we consider the I sub-sequences yielding the largest log-likelihoods $s_{n,i}$, which end at positions $q_{n,i}$. Let $d_n = q_{n,1}$ and $d'_n = |\boldsymbol{y}_n| - q_{n,1}$ be the distance of the sub-sequence with the largest log-likelihood to the 3' and 5' end of the UTR, respectively. Let $(p_{n,1}, \ldots, p_{n,I})$ denote the positions $(q_{n,1}, \ldots, q_{n,I})$ sorted ascendingly. Let $\boldsymbol{z}_n = (s_{n,1}, \ldots, s_{n,I}, d_n, d'_n, p_{n,1}, \ldots, p_{n,I})$ denote the vector of these features representing UTR $\boldsymbol{y}_n$.

By inspecting histograms of the scores $s_{n,i}$, we find that these may be modeled by a mixture of two Gaussian densities, i.e.,

$$P(s_{n,i}|\boldsymbol{\beta}^s_{c,i}) = P(u^s = 1|\boldsymbol{\beta}^{s,m}_{c,i})\, \mathcal{N}(s_i|\mu_{1,i,c}, \kappa_{1,i,c}) + P(u^s = 2|\boldsymbol{\beta}^{s,m}_{c,i})\, \mathcal{N}(s_i|\mu_{2,i,c}, \kappa_{2,i,c}),$$

where $\boldsymbol{\beta}^s_{c,i} = (\boldsymbol{\beta}^{s,m}_{c,i}, \mu_{1,i,c}, \kappa_{1,i,c}, \mu_{2,i,c}, \kappa_{2,i,c})$, $\mu_{k,i,c}$ and $\kappa_{k,i,c}$ denote the mean and the log-precision of Gaussian density k, respectively, and the component probabilities $P(u^s = u|\boldsymbol{\beta}^{s,m}_{c,i})$ are parameterized in analogy to equation (1).
To allow for variability in TS positioning, we model d_n and d'_n each by a mixture of two gamma densities, i.e.,

$$P(d_n|\boldsymbol{\beta}^d_c) = P(u^d = 1|\boldsymbol{\beta}^{d,m}_c)\, \mathcal{G}(d_n|\alpha^d_{1,c}, \beta^d_{1,c}) + P(u^d = 2|\boldsymbol{\beta}^{d,m}_{c,i})\, \mathcal{G}(d_n|\alpha^d_{2,c}, \beta^d_{2,c}),$$

where $\boldsymbol{\beta}^d_c = (\boldsymbol{\beta}^{d,m}_c, \alpha^d_{1,c}, \beta^d_{1,c}, \alpha^d_{2,c}, \beta^d_{2,c})$, and $\alpha^d_{k,c}$ and $\beta^d_{k,c}$ denote the log-shape and log-rate of gamma density k, respectively. We define the density $P(d'_n|\boldsymbol{\beta}^{d'}_c)$ in analogy.

We model the distances $p_{n,i+1} - p_{n,i}$ by another gamma density, i.e.,

$$P(p_{n,i+1} - p_{n,i}|\boldsymbol{\beta}_c^p) = \mathcal{G}(p_{n,i+1} - p_{n,i}|\alpha_c^p, \beta_c^p),$$

where $\boldsymbol{\beta}_c^p = (\alpha_c^p, \beta_c^p)$.

The complete likelihood of $\boldsymbol{z}_n$ representing UTR $\boldsymbol{y}_n$ of gene n employing convenient independence assumptions amounts to

$$P(\boldsymbol{z}_n|c, \boldsymbol{\beta}_c) \propto \prod_{i=1}^{I} P(s_{n,i}|\boldsymbol{\beta}_{c,i}^s)\ P(d_n|\boldsymbol{\beta}_c^d)\ P(d_n'|\boldsymbol{\beta}_c^{d'}) \prod_{i=1}^{I-1} P(p_{n,i+1} - p_{n,i}|\boldsymbol{\beta}_c^p). \quad (8)$$

In the following studies, we use $I = 5$.

In analogy to equation (6), we define the class posterior in terms of likelihoods $P(\boldsymbol{z}_n|c, \boldsymbol{\beta}_c)$ and a-priori class probabilities $P(c|\boldsymbol{\beta})$. As for the training of the TS model, we optimize the parameters with respect to the discriminative MSP principle (cf. equation (7)) using a training data set of target and non-target genes. In this case, we use beta priors on the parameters of the component probabilities, normal-gamma priors on the parameters of the Gaussian densities, and the conjugate prior according to the definition of the exponential family for the gamma densities. Again, we use an ESS of 4 for both classes. We finally predict target genes based on the class posterior.

3 Results & Discussion

In the following, we first investigate if CoProHMMs can learn characteristics of TSs from data. To this end, we use TSs predicted by existing approaches. Second, we evaluate the utility of CoProHMMs for the prediction of target genes of miRNAs on benchmark data.

3.1 Pilot study: Learning CoProHMMs from predictions

We learn CoProHMMs on the predictions of miRanda and TargetScan to investigate if CoProHMMs can learn the rules implemented into these approaches from their predictions. We choose miRanda and TargetScan, because their approaches differ notably. If CoProHMMs can detect such characteristics from predictions, we might expect that they are also capable of learning novel or refined rules of miRNA-TS binding from experimentally verified TS.

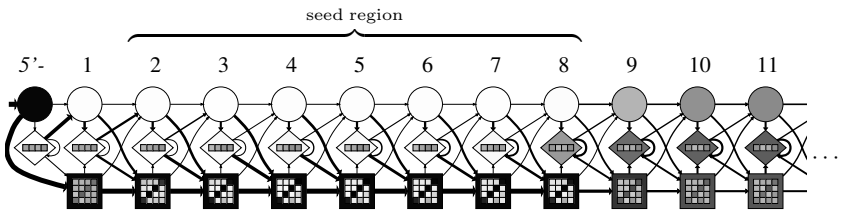
We extract all human TSs and associated miRNAs predicted by TargetScan and miRanda from miRNAMap[1] [HCT+08]. For TargetScan, we use all 244,389 TSs, while we randomly sample 500,000 TSs from the predictions of miRanda. We generate a non-target data set by randomly selecting miRNAs from the mature human miRNAs listed at miRBase[2] [GJSvDE08]. As non-TSs of these miRNAs, we randomly draw 500,000 sub-

[1] ftp://mirnamap.mbc.nctu.edu.tw/miRNAMap2/miRNA_Targets/Homo_sapiens/miRNA_targets_hsa.txt.tar.gz

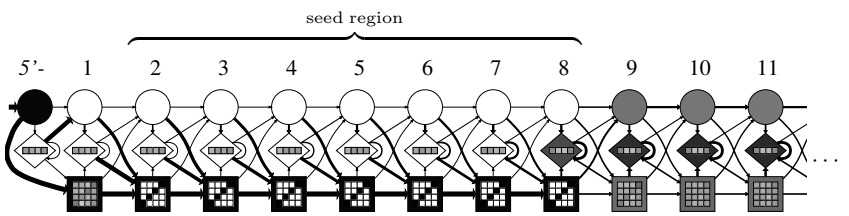
[2] http://www.mirbase.org

sequences of length $|\boldsymbol{r}| \pm 3$ from 3'-UTRs of human genes according to NCBI Genbank[3] human genome build 37.1.

We present a graphical representation of the CoProHMMs learned on the miRanda data set and the TargetScan data set in Fig. 2. Here, we depict only the most interesting region around the seed, while the complete CoProHMMs for miRanda and TargetScan as well as other approaches are available online[4]. For the states, we use the same shapes as in Fig. 1. The thickness of outgoing edges represents the transition probabilities to the successors of a node. We illustrate the emission probabilities of insert states by a row of grayscale boxes, where the first box corresponds to A, the second box corresponds to C, the third box corresponds to G, and the fourth box corresponds to U. The darker a box, the higher is the corresponding emission probability. In analogy, the conditional emission probabilities of match states are represented by a matrix comprising such rows, where each row corresponds to the conditional probability distribution given one nucleotide of the miRNA. The probabilities of visiting a state are visualized by the darkness of the background of each node. The darker the background of a node the higher the probability of visiting this node.



(a) miRanda data set



(b) TargetScan data set

Figure 2: CoProHMMs learned on the miRanda data set (a) and TargetScan data set (b).

Considering the CoProHMM learned on the miRanda data set, we recover many rules built into miRanda. From the conditional emission probabilities of the match states, we observe a general tendency to complementary base pairings between the TS and the miRNA. This tendency is especially pronounced for the match states in the seed region, but can also be observed for the match states at position 1 and positions 9 to 11. We also detect a slight preference for `G:U` wobble basepairs. These observations are most likely a result of the Smith-Waterman like alignment employed by miRanda. Additionally, miRanda assigns a

[3] `http://www.ncbi.nlm.nih.gov`
[4] `http://www.jstacs.de/index.php/MiRNAs`

weight of 2 to the first 11 positions of the alignment, which is reflected by the increased probabilities of visiting match states in the seed region, although this preference already begins to decline at position 8 of the learned CoProHMM.

As a second example, we consider the CoProHMM learned on the TargetScan data set in Fig. 2(b). Notable differences between the CoProHMM for the TargetScan data set and the miRanda data set can be observed for the conditional emission probabilities at the match states. At positions 2 to 8 of Fig. 2(b), we find complementary basepairs almost exclusively, while a slight preference for complementary basepairs is present at the bordering positions 1 and 9. In contrast, the remaining positions exhibit only very slight preferences for specific basepairs. Again, these findings are closely related to the main characteristics built into TargetScan. The perfect complementarity at positions 2 to 8 of the CoProHMM reflects the requirements of TargetScan. We also observe a preference for complementary basepairs at positions 1 and 9, which most likely can be attributed to the fact that initial perfect matches in the seed region may be elongated to either side in TargetScan.

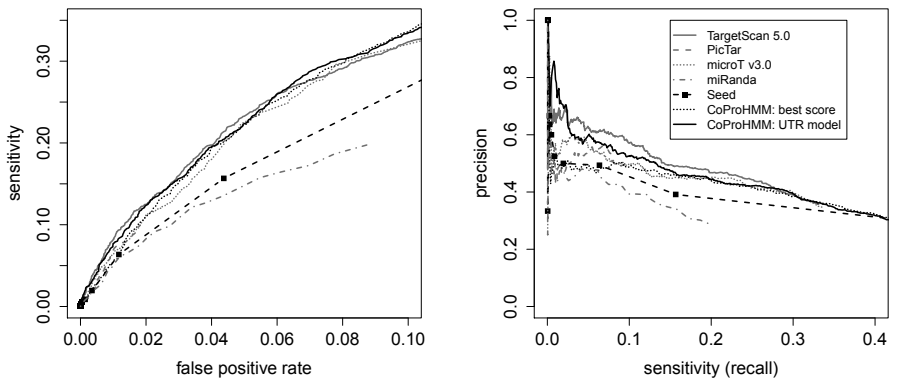
These findings suggest that CoProHMMs are indeed capable of recovering the rules built into miRanda and TargetScan from prediction and, hence, may also be capable of inferring the rules underlying miRNA-TS binding from experimentally verified TSs, once these become available in sufficient quantity.

3.2 Benchmark study: Predicting miRNA target genes

We investigate the utility of CoProHMMs for the prediction of miRNA target genes using the pSILAC data of Selbach *et al.*, which have also been used in recent benchmark studies [SST$^+$08, AMP$^+$09]. To this end, we learn a CoProHMM using a foreground data set that comprises 12 verified TSs and 667 predicted TSs within UTRs of verified target genes extracted from mirecords[5] v. 1 [XZC$^+$09]. As these TSs are too few to reliably learn the models, we also include the TargetScan data set and 405,569 TSs predicted by DIANA-microT. We use predictions of these two approaches, since they yield reasonable precisions in the benchmark studies. We use the same background data set as in the pilot study. We assign a weight of 500 to all verified TSs and a weight of 50 to all predicted TSs in verified target genes to reflect our increased confidence in these data, while we assign a weight of 1 to all other TSs. All TSs of miRNAs contained in the Selbach benchmark data set are excluded when training the CoProHMM to allow for unbiased evaluation.

We extract the UTRs of all genes considered in [SST$^+$08] according to [AMP$^+$09]. For these genes, Selbach *et al.* measured the influence of overexpression or underexpression of a miRNA on the abundance of the corresponding proteins for 5 different miRNAs. For each of these miRNAs, we partition the UTRs into target and non-target UTRs using a threshold of -0.2 on the protein log-fold changes. We assess the performance of the UTR model using the predictions of the CoProHMM in a 5-fold cross validation. In each iteration of the cross validation, we train the parameters of the UTR model on the numeric vectors $\boldsymbol{z}_n$ obtained for 4 of the 5 miRNAs, and we compute the log-likelihood ratios using this trained UTR model for the numeric vectors obtained for the remaining miRNA.

[5] `http://mirecords.biolead.org/download_data.php?v=1`



(a) ROC curve

(b) Precision-recall curve

Figure 3: ROC curve (a) and precision-recall curve (b) of the classifier using the UTR model (solid black line) and the classifier using the best score of the CoProHMM within each UTR sequence (dotted black line) compared to other approaches.

In analogy to [AMP+09], we finally use all log-likelihood ratios to compute sensitivity, precision, and false positive rate for different thresholds.

In Fig. 3, we compare the performance of the classifier using the UTR model (solid black line) to other approaches by means of the precision-recall curve and the ROC curve. As a reference, we also include the performance of a classifier that only uses the best score of the CoProHMM over each UTR sequence, i.e., $s_{n,1}$, (dotted black line). Considering Fig. 3(a), we find that even this classifier using only the best score yields a substantially higher sensitivity than miRanda and Seed for a broad range of false positive rates. Surprisingly, the classifier using the simple UTR model, which does not exploit conservation across species, achieves comparable or slightly improved sensitivities compared to miRanda, Seed, PicTar, and microT, while it performs only slightly worse than TargetScan 5.0 for false positive rates below 0.06.

Turning to the precision-recall curve in Fig. 3(b), we find a similar picture. Notably, the classifier using the UTR model again achieves comparable or even higher precisions than miRanda, Seed, PicTar, and microT. However, it can outperform TargetScan 5.0 only for very low sensitivities and yields lower precisions for sensitivities between 0.03 and 0.28.

The performance of both classifiers using CoProHMMs is astonishing, because, in contrast to most of the other approaches, they do not exploit conservation across different species. Hence, the inclusion of cross-species conservation into CoProHMMs and the proposed UTR model, and the integration of CoProHMMs into other approaches might be a worthwhile direction of future research.

4 Conclusions

miRNAs are involved in the regulation of many cellular processes, and the prediction of miRNA targets is one of the most active fields of bioinformatics. Here, we propose a novel statistical model called conditional profile HMM (CoProHMM) for learning the rules of miRNA-TS interaction from data. We demonstrate that CoProHMMs are capable of reconstructing patterns of miRNA-TS binding built into existing programs from predictions of these approaches.

Conservation is key feature of most miRNA target prediction approaches leading to higher precision at the expense of sensitivity. Interestingly, we find in a benchmark study that a simple UTR model utilizing CoProHMMs yields a competitive precision compared to leading approaches for predicting target genes, although it does not exploit conservation across species.

We anticipate that the number of experimentally verified TSs will rapidly increase in the next years. Only recently, [CZMD09, HLB+10] have independently published novel biological data that shed light on miRNA targeting. Briefly, the two experimental approaches use in-vivo crosslinking, Ago2 immunoprecipitation and cDNA sequencing, and have been able to determine TSs of several miRNAs with high accuracy. Since the power of statistical approaches like CoProHMMs highly depends on the quality of the training data, we might speculate that the performance of CoProHMMs will even increase using these data. Additionally, CoProHMMs might be a suitable approach to extract new and refined rules of miRNA-TS binding from such verified TSs.

We make an implementation of CoProHMMs and the UTR model available to the scientific community with the next release of the open source Java library Jstacs[6].

References

[AMP+09] Panagiotis Alexiou, Manolis Maragkakis, Giorgos L. Papadopoulos, Martin Reczko, and Artemis G. Hatzigeorgiou. Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, 25(23):3049–3055, 2009.

[BB01] Pierre Baldi and Søren Brunak. *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge, London, 2nd edition, 2001.

[BSRC05] Julius Brennecke, Alexander Stark, Robert B. Russell, and Stephen M. Cohen. Principles of MicroRNA–Target Recognition. *PLoS Biology*, 3(3), 2005.

[CdM05] Jesús Cerquides and Ramon López de Mántaras. Robust Bayesian Linear Classifier Ensembles. In *Proceedings of the 16th European Conference on Machine Learning*, volume 3720 of *Lecture Notes in Computer Science*, pages 72–83. Springer, 2005.

[CZMD09] Sung Wook Chi, Julie B. Zang, Aldo Mele, and Robert B. Darnell. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254):479–486, 07 2009.

[EJG+03] Anton Enright, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora Marks. MicroRNA targets in Drosophila. *Genome Biology*, 5(1):R1, 2003.

[6] http://www.jstacs.de

[FFBB09] Robin C. Friedman, Kyle Kai-How Farh, Christopher B. Burge, and David P. Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105, 2009.

[GJSvDE08] Sam Griffiths-Jones, Harpreet Kaur Saini, Stijn van Dongen, and Anton J. Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Research*, 36(suppl_1):D154–158, 2008.

[GKK+07] Jan Grau, Jens Keilwagen, Alexander Kel, Ivo Grosse, and Stefan Posch. Supervised posteriors for DNA-motif classification. In Claudia Falter, Alexander Schliep, Joachim Selbig, Martin Vingron, and Dirk Walther, editors, *German Conference on Bioinformatics*, volume 115 of *Lecture Notes in Informatics (LNI) - Proceedings*, Bonn, 2007. Gesellschaft für Informatik.

[HCT+08] Sheng-Da Hsu, Chia-Huei Chu, Ann-Ping Tsou, Shu-Jen Chen, Hua-Chien Chen, Paul Wei-Che Hsu, Yung-Hao Wong, Yi-Hsuan Chen, Gian-Hung Chen, and Hsien-Da Huang. miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes. *Nucleic Acids Research*, 36(suppl_1):D165–169, 2008.

[HGC95] David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. In *Machine Learning*, pages 197–243, 1995.

[HLB+10] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S. Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. 141(1):129–141, 04 2010.

[KBM+94] Anders Krogh, Michael Brown, I. Saira Mian, Kimmen Sjölander, and David Haussler. Hidden Markov Models in Computational Biology : Applications to Protein Modeling. *Journal of Molecular Biology*, 235(5):1501 – 1531, 1994.

[KGP+05] Azra Krek, Dominic Grun, Matthew N Poy, Rachel Wolf, Lauren Rosenberg, Eric J Epstein, Philip MacMenamin, Isabelle da Piedade, Kristin C Gunsalus, Markus Stoffel, and Nikolaus Rajewsky. Combinatorial microRNA target predictions. *Nature Genetics*, 37(5):495–500, 05 2005.

[LBB05] Benjamin P. Lewis, Christopher B. Burge, and David P. Bartel. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell*, 120(1):15 – 20, 2005.

[LSJR+03] Benjamin P. Lewis, I-hung Shih, Matthew W. Jones-Rhoades, David P. Bartel, and Christopher B. Burge. Prediction of Mammalian MicroRNA Targets. *Cell*, 115(7):787 – 798, 2003.

[Mac98] David J. C. MacKay. Choice of Basis for Laplace Approximation. *Machine Learning*, 33(1):77–86, 1998.

[MRS+09] M. Maragkakis, M. Reczko, V. A. Simossis, P. Alexiou, G. L. Papadopoulos, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas, and A. G. Hatzigeorgiou. DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Research*, 37(suppl_2):W273–276, 2009.

[SST+08] Matthias Selbach, Bjorn Schwanhausser, Nadine Thierfelder, Zhuo Fang, Raya Khanin, and Nikolaus Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63, 09 2008.

[XZC+09] Feifei Xiao, Zhixiang Zuo, Guoshuai Cai, Shuli Kang, Xiaolian Gao, and Tongbin Li. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Research*, 37(suppl_1):D105–110, 2009.