BMC Bioinformatics

BioMed Central

Research

HHMMiR: efficient *de novo* prediction of microRNAs using hierarchical hidden Markov models

Sabah Kadri*[1], Veronica Hinman[2] and Panayiotis V Benos*[3]

Address: [1]Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA, [2]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA and [3]Department of Computational Biology, University of Pittsburgh, Pittsburgh, PA 15260, USA

Email: Sabah Kadri* - sskadri@andrew.cmu.edu; Veronica Hinman - vhinman@cmu.edu; Panayiotis V Benos* - benos@pitt.edu

* Corresponding authors

from The Seventh Asia Pacific Bioinformatics Conference (APBC 2009)
Beijing, China. 13–16 January 2009

Published: 30 January 2009

BMC Bioinformatics 2009, **10**(Suppl 1):S35 doi:10.1186/1471-2105-10-S1-S35

This article is available from: http://www.biomedcentral.com/1471-2105/10/S1/S35

Abstract

Background: *MicroRNA*s (miRNAs) are small non-coding single-stranded RNAs (20–23 nts) that are known to act as post-transcriptional and translational regulators of gene expression. Although, they were initially overlooked, their role in many important biological processes, such as development, cell differentiation, and cancer has been established in recent times. In spite of their biological significance, the identification of miRNA genes in newly sequenced organisms is still based, to a large degree, on extensive use of evolutionary conservation, which is not always available.

Results: We have developed HHMMiR, a novel approach for *de novo* miRNA hairpin prediction in the absence of evolutionary conservation. Our method implements a *Hierarchical Hidden Markov Model* (HHMM) that utilizes region-based structural as well as sequence information of miRNA precursors. We first established a template for the structure of a typical miRNA hairpin by summarizing data from publicly available databases. We then used this template to develop the HHMM topology.

Conclusion: Our algorithm achieved average sensitivity of 84% and specificity of 88%, on 10-fold cross-validation of human miRNA precursor data. We also show that this model, trained on human sequences, works well on hairpins from other vertebrate as well as invertebrate species. Furthermore, the human trained model was able to correctly classify ~97% of plant miRNA precursors. The success of this approach in such a diverse set of species indicates that sequence conservation is not necessary for miRNA prediction. This may lead to efficient prediction of miRNA genes in virtually any organism.

Background

MicroRNAs

*MicroRNA*s (miRNAs) are small (~22 nucleotide long) non-coding RNAs that are part of a eukaryote-specific system of gene regulation at the RNA level. MiRNAs act as post-transcriptional regulators of gene expression by base pairing with their target mRNAs. MiRNAs are primarily transcribed by *RNA Pol II* [1] as regions of longer RNA molecules (pri-miRNA) [2]. Individual pre-miRNA loops (~70 nts) are cleaved from the pri-miRNA by RNAse III enzyme, *Drosha* and transported into the cytoplasm by *RAN-GTP* and *Exportin 5* [3] to be processed further to a ~22 nt long duplex, with 3' overhangs, by *Dicer* [4,5]. This duplex is commonly referred to as the miRNA:miRNA* duplex, where miRNA* is complementary to the miRNA. The miRNA:miRNA* duplex is subsequently unwound and the mature miRNA is loaded into multi-protein RISC (RNA-induced silencing complex) [6] while miRNA* usually degrades. In some cases, both miRNA and miRNA* are functional [7]. The miRNA biogenesis is illustrated in Figure 1. Mature miRNAs can cause translation inhibition or mRNA cleavage, depending on the degree of complementarity between the miRNA and its target sequence. Each miRNA can have multiple targets and each gene can be targeted by multiple miRNAs. It has been predicted that more than one third of human genes is regulated by miRNAs [8].

Plant and animal miRNAs differ not only in their biogenesis, but also in target-miRNA interactions. Plant miRNAs base pair with their targets with perfect or near-perfect complementarity and they regulate their targets mostly through mRNA cleavage at single sites in coding regions. Animal miRNAs usually base pair with 3' UTRs of the mRNAs at multiple target sites through imperfect complementarity. Due to these and other differences, it has been suggested that this regulation mechanism may have evolved independently in plants and animals [9]. Some viruses have also been shown to encode miRNAs that play a role in expression regulation of host genes [10].

MiRNA identification

The first animal miRNA genes, *let-7* and *lin-4*, were discovered in *Caenorhabditis elegans* by forward genetics [11-13]. Currently, miRNA genes are biochemically identified by cloning and sequencing size-fractionated cDNA libraries. The main limitation of this method is that lowly expressed miRNAs may be missed [14]. Although deep sequencing can help overcome this problem, this is currently a costly solution. Still, some miRNAs may be difficult to clone due to their sequence composition and possible post-transcriptional modifications [14-16]. Deep sequencing is being used on a large scale to identify small non-coding RNAs, but this is an expensive method and can only identify miRNAs expressed in a single cell type or in a given condition.

Computational predictive methods are fast and inexpensive and a number of approaches have been developed to predict miRNA genes, genome-wide. However, most of these approaches depend heavily on conservation of hairpins in closely related species [17-20]. Some methods have used clustering or profiling to identify miRNAs, [17,21,22]. The approach of Bentwich *et al.* [23] is interesting in that the whole genome is folded and scores are assigned to hairpins based on various features, including hairpin structural features and folding stability.

Machine learning approaches in the past have used support vector machines with high dimensional basis functions for classification of genomics hairpins [22,24,25]. Some of these methods depend on cross-species conservation for classification, while others do motif finding using multiple alignments. More recently, HMMs have been used in modelling miRNAs using both, evolutionary information and features related to the secondary structure [26].

Hierarchical Hidden Markov Models

Hierarchical Hidden Markov Models (HHMMs) constitute a generalization of Hidden Markov Models (HMMs). They have been successfully used for modelling stochastic levels and length scales [27]. In biology, HHMMs have been used in the past to model vertebrate splice sites [28] and more recently in modelling *cis*-regulatory modules [29]. An HHMM has two types of states: *internal states* and *production states*. Each internal state has its own HHMM but cannot emit symbols by itself. It can activate a sub-state by a vertical transition. Sub-states can also make vertical transitions, until the lowest level in the hierarchy (production state) is reached. Production states are the only states that can emit symbols from the alphabet *via* their own probability distributions. Sub-states at the same level of hierarchy will be activated through horizontal transitions till an "end state" is reached. Every level has only one "end state" for each parent state that shifts control back to the parent. Thus, each internal state can emit sequences instead of single symbols. The node at the highest level of the hierarchy is called the "root" node while the leaf nodes are the productions states. Please refer to *Methods* for information about HHMM parameters and their estimation.

In this article, we report the results on the performance of an HHMM we developed for modelling miRNA hairpins. Although the model was trained on human sequences only, it was able to classify accurately hairpins from species as distant as worm, flies and plants, indicating that the degree of sequence and structural conservation for these genes may be high.

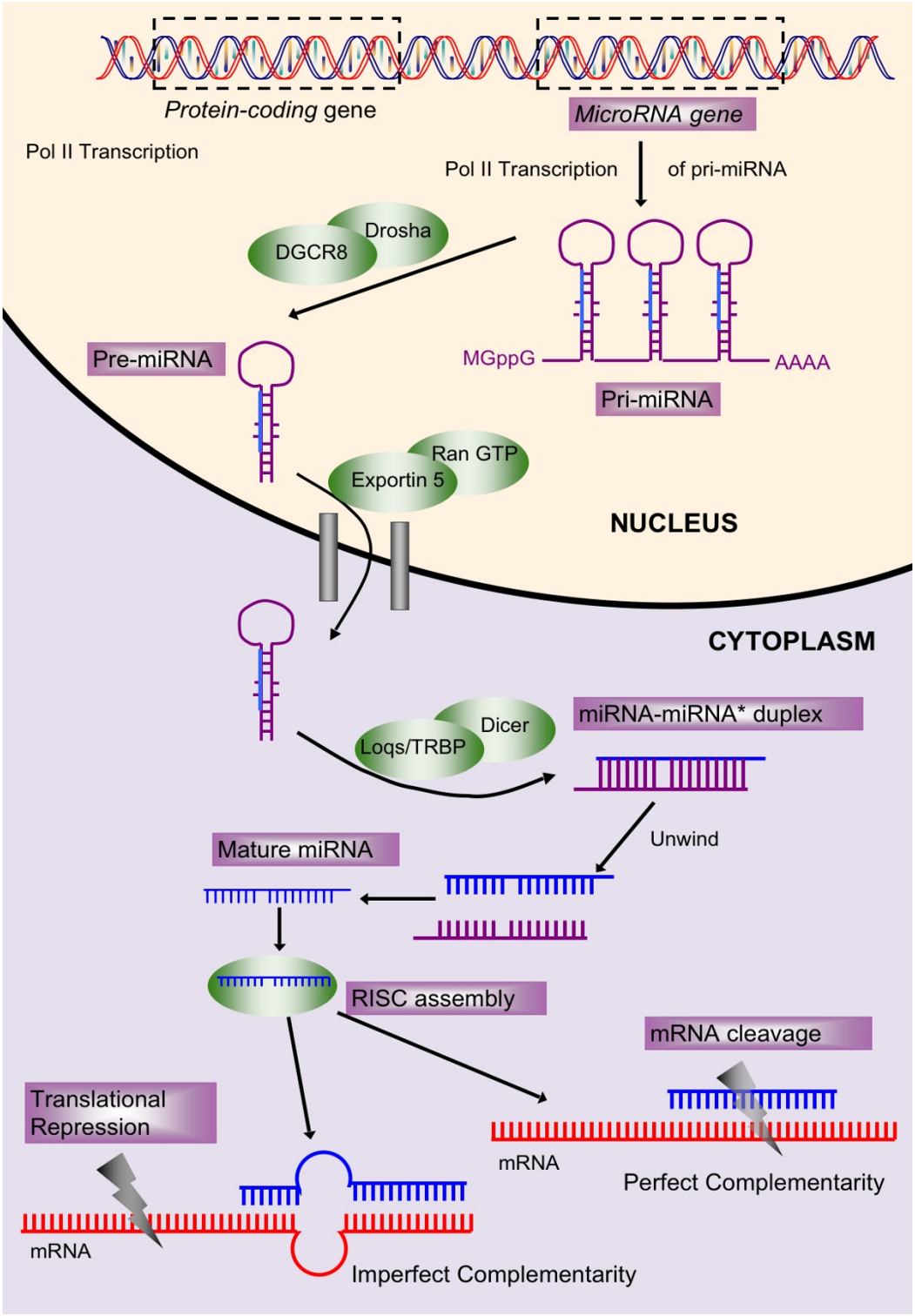

Figure 1

Biogenesis of microRNAs. miRNA genes are transcribed in the nucleus, where they undergo processing by DGCR8/Pasha and the RNAse III family enzyme, Drosha. The pre-miRNA is then transported into the cytoplasm where it is processed by Dicer, and the cofactor TRBP to generate a ~22 nt miRNA:miRNA* duplex. After unwinding, the miRNA forms part of the RISC assembly and causes mRNA degradation or translational repression.

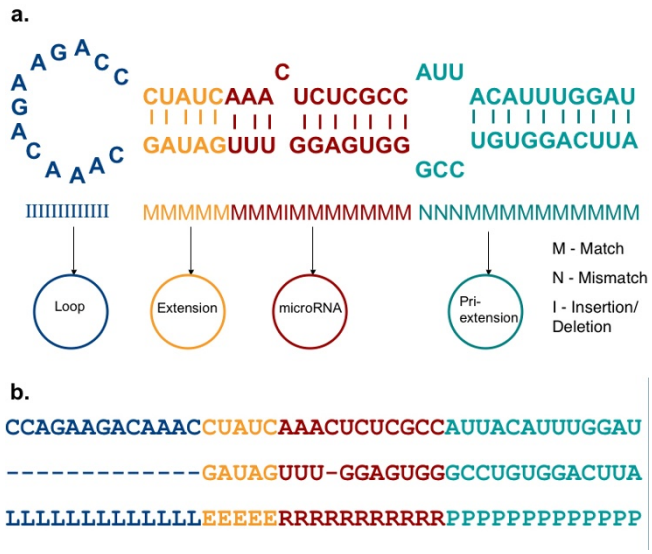

Figure 2
The miRNA hairpin. **(a)** *Template*: In our model, the miRNA precursor has four regions- "Loop" is the bulge and the *loop* state outputs *indels* only; "Extension" is a variable length region between the miRNA duplex and the loop; "microRNA" represents the duplex, without 3' overhangs; "Pri-extension" is the rest of the hairpin. The latter three states can output *matches*, *mismatches* and *indels*. (The nucleotides distribution and lengths are not to scale) **(b)** *Labelled precursor*: The precursor shown in (a) is labelled according to the regions it represents. This is the input format of training data for HHMMiR. L: Loop; E: Extension; R: MiRNA; P: Pri-miRNA.

Results

Data summarization

We consider the hairpin stem-loop for predictions since it is structurally, the most prominent feature during biogenesis (Figure 1). MiRNA genes can be divided into four regions depicted in Figure 2a. After transcription, the RNA strand folds to form the hairpin precursor (Figure 1 and Figure 2a). The "loop" is the bulged end of the hairpin. The "miRNA" region defines the miRNA-miRNA* duplex (sans the 3' overhangs) that is processed by Dicer and further unwound. The region of the precursor extending from the end of the loop to the "miRNA" region is called the "extension". This region can be of variable length. The part of the hairpin sequence beyond the "miRNA" region may be part of the pri-miRNA in the nucleus and processed by Drosha. Thus, it has been named as "pri-extension", as suggested in Saetrom *et al.* [30].

The results presented in Table 1 show that the differences that exist between vertebrate and invertebrate miRNA genes are rather small. So, a probabilistic method trained in data from one organism is likely to perform well in another organism. As evident from the results in Table 1, the differences between length distributions of plant and animal precursors are relatively drastic, with the former having longer extension regions. The lengths of miRNAs and loops, however, are conserved across the two kingdoms. More information about species-specific differences is provided in Additional File 1. These genomes constitute an excellent test set for our algorithm in that they span various taxonomic groups, with different miRNA characteristics. Thus, it will be very useful to see how well an HHMM trained on (say) human sequences will be able to predict miRNA stem-loops in another vertebrate or invertebrate species and plants.

HHMM model

HHMMiR is built around the miRNA precursor template illustrated in Figure 2a. The figure presents the four characteristic regions of stem-loop of a typical miRNA gene as described above. The length distributions of each of these regions are derived from Table 1. Each region, except the loop itself has three states: *match*, *mismatch*, and *insertion/deletion* (*indel*). *Match* means a base pairing at that position in the stem-loop, while *mismatch* means bulges on both arms at that position in the folded hairpin.*Indel* means that a base in one strand has no counterpart in the opposite strand. The loop will only have the *indel* state. Examples of these states are presented in Figure 2a.

The HHMM resulting from this scheme has three levels (Figure 3). *Hairpin* is the root node and can vertically transition to its *Loop* substate only. In our model, every hairpin begins with a loop. The four internal states at the

Table 1: Characteristics of miRNA hairpins in various taxonomic groups.

	HP	LP	MIR	EXT	PRI
Mean					
Vertebrates	86.7	7.3	22.0	5.0	12.6
Invertebrates	91.8	7.9	22.2	5.8	13.8
Plants	119.5	6.8	21.3	22.8	12.5
Std. Dev.					
Vertebrates	13.8	3.5	0.9	3.4	7.0
Invertebrates	13.1	3.9	1.3	4.5	5.9
Plants	43.2	3.7	1.0	18.5	9.9
Minimum					
Vertebrates	55	3	16	0	0
Invertebrates	54	3	18	0	0
Plants	57	3	16	0	0
Maximum					
Vertebrates	153	22	26	34	50
Invertebrates	215	30	28	55	32
Plants	337	35	24	102	78

HP: Hairpin length; *LP*: Loop length; *MIR*: MiRNA length; *EXT*: Distance of miRNA duplex from end of loop; *PRI*: Length of extension from end of miRNA to end of precursor. The list of organisms used for this Table is provided as *Supplementary Data*.

second level correspond to the four main regions of the hairpin from Figure 2a. This level also has an *End* (L_{end}) state to transfer control back to the *Hairpin*. Each internal state has a probabilistic model at the next lower level. A *Loop* cannot have base pairs and thus, has only one substate: *I* (*Indel)*. The *Extension* state can only emit an *M* (*match*) state, when entered, since a mismatch or indel would become part of the loop. The *miRNA* and *pri-Ext* states can begin with a match, mismatch or indel. Each of these states has an *End* state (L_{end}, R_{end}, P_{end} respectively)(see Figure 3).

Datasets and alphabet selection

The training dataset contained a total 527 human miRNA precursors (positive dataset) and ∼500 random hairpins (negative dataset), based on criteria derived from summarization (see *Methods*). The *RNAfold* program from Vienna Package [31] was used to obtain the secondary structure of these hairpins with the *minimum fold energy* (*mfe*). The parameters of the model were estimated using a modified Baum-Welch algorithm (see *Methods* for details on data sets and algorithms). All tests were conducted with 10-fold cross validation with random sampling.

We tested our model on two alphabets: Σ_1 with *matches M* = {AU, GC, GU}, *indels I* = {A-, G-, C-, U-} and *mismatches* N_1 = {AA, GG, CC, UU, AC, AG, CU}; and Σ_2, which is similar to Σ_1 except that the mismatch set is more concise: N_2 = {XX, XY}, where XX stands for one of {AA, GG, CC, UU} and XY stands for one of {AC, AG, CU}. In our alphabet, a *match*, say, AU has the same probability as UA, that is, an 'A' on either stem base paired with 'U' on the other stem. Cross-validation tests using *Maximum Likelihood Estimate* (MLE) showed that the model with alphabet Σ_1 performed substantially better, both in terms of sensitivity and specificity (Table 2) (see *Methods* for more details on these calculations).

It is surprising that Σ_1 performs better than Σ_2, because one would expect that mismatches in the stem-loop would not be characteristic of the miRNA sequence, since they do not contribute to the base pairing of the stem and

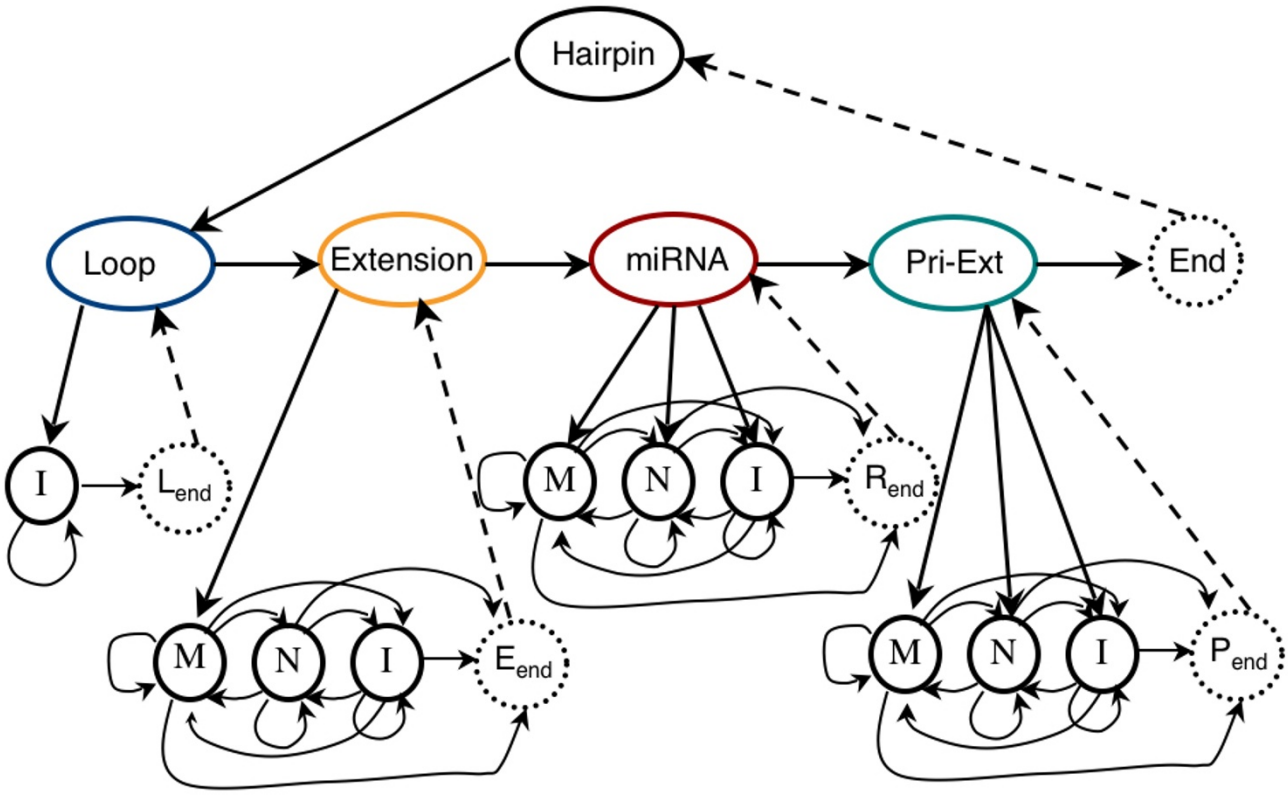

Figure 3
The HHMM state model (based on the microRNA hairpin template). The oval shaped nodes represent the *internal states*. The colours correspond to the biological region presented in Figure 2a. The circular solid lined nodes correspond to the production states. The dotted lined states correspond to the silent end states. M: *Match* states, N: *Mismatch* states, I: *Indel* states, L_{end}: Loop end state, R_{end}: miRNA end state, P_{end}: pri-extension end state.

Table 2: Results for different alphabet sizes: Σ_1 (larger alphabet) shows better accuracy than Σ_2 (smaller alphabet)

Alphabet	Sn	Sp	FDR
Σ_1	74.5	94.1	15.8
Σ_2	55.0	48.5	51.0

Sn: Sensitivity; Sp: Specificity; FPR: False Positive rate; FDR: False Discovery rate. All numbers are in percentages.

thus the overall folding energy, on which other algorithms are based [23]. Furthermore, Σ_1 alphabet has more parameters. In order to rule out that the better performance is due to parameter overfitting, we repeated training with multiple datasets of different sizes and the results remained the same (*data not shown*). In the remaining of this paper we use the Σ_1 alphabet.

Training algorithms: performance evaluation

We implemented and compared variations of two existing algorithms for parameter estimation: Baum-Welch and MLE. The positive model was trained using MLE since by nature the positive training data (stem-loop hairpins) can be labelled as *loop*, *extension*, *miRNA* and *pri-extension* (Figure 2b) using existing annotations. Negative data on the other hand, are obviously unlabelled, so both algorithms were compared for training with this dataset. We will call the MLE trained negative model, MLE-HHMMiR, whereas the Baum Welch trained model will be called BW-HHMMiR for this evaluation. For MLE-HHMMiR, we used length distributions from database summarization (Table 1) to perform *random labelling* of the four regions on the negative datasets. Overall, we found both methods to perform practically the same. The area under the ROC curve (Figure 4) for the MLE-HHMMiR is 0.912 whereas for BW-HHMMiR is 0.920. The ratio of the log-likelihoods output by the two models decides the fate of the test hairpin. In order to decide a threshold for this ratio, the trade-off between sensitivity and specificity was considered by calculating the *Mathews correlation coefficient* (Table 3). The highest Mathews correlation coefficient value was 0.73 for BW-HHMMiR and 0.71 for MLE-HHMMiR, corresponding to likelihood ratio thresholds of 0.71 and 0.99, respectively. BW-HHMMiR achieved an average 84% sensitivity and 88% specificity using the 0.71 ratio as thresholds. Even though, the difference between the performances of the two algorithms is not great, we choose BW-HHMMiR for further tests. This is because MLE-HHMMiR depends on *random labelling* of hairpins and thus, performance will vary according to the labelling. The drawback of the Baum-Welch method is that it might be trapped on local optima, depending on the initialization. This problem is sometimes addressed by running the algorithm multiple times with different starting points. We use a uniform distribution for this initialization but can also use background frequencies for the same by folding the entire genome in question and then performing hairpin extraction for the same. In order to account for the absence of certain base pairs or *indels* in a certain sequence while using Baum-Welch, we introduce pseudo-counts to correct for the same.

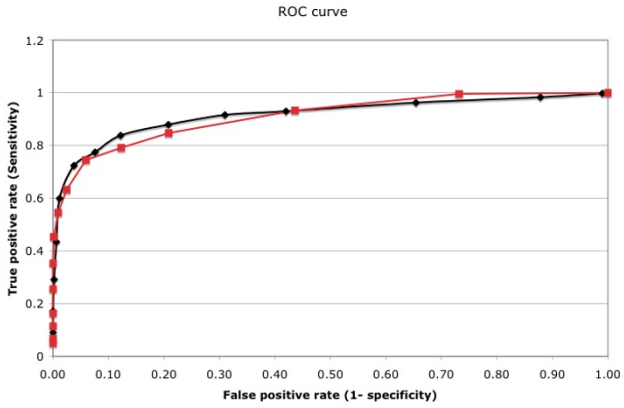

Figure 4
ROC curves for Baum-Welch and MLE training on the negative model. 10-fold cross-validation used with Baum-Welch (*black curve*) and MLE (*red curve*) for training the negative model. Positive model was trained using MLE in both cases.

Table 3: Results for cross-validation using different algorithms.

Method	Sn (SD)	Sp (SD)	MCC	FDR (SD)
Baum-Welch	84.0 (18.6)	88.0 (6.6)	0.73	11.8 (5.6)
MLE	74.5 (13.7)	94.1 (2.7)	0.71	15.9 (8.0)

Sn: sensitivity; Sp: specificity; MCC: Mathew's correlation coefficient; FDR: False Discovery Rate. Sn, Sp and FDR report the average percent values; standard deviations are reported in parentheses.

Testing prediction efficiency in other organisms

Next, we examined how well our model trained on human sequences could predict known miRNAs in other species. In particular, HHMMiR was tested on the following species: *M. musculus* (mammals), *G. gallus* (birds), *D. rerio* (fish), *C. elegans* (worms), *D. melanogaster* (flies), *A. thaliana* and *O. sativa*(plants). These species were chosen as representatives of their respective taxonomic groups, and because they are well studied and annotated. The results are shown in Table 4. HHMMiR is able to predict 85% of most animal precursors. Its overall sensitivity was also about 85%. What is more surprising, however, is the higher performance we observe in prediction of plant precursors, given the differences in length distributions of the miRNA stem-loops between plants and animals (Table 1).

Table 4: Results of tests on other species.

Organism	Total hairpins	% correctly predicted
M. musculus	422	74.7
G. gallus	147	89.1
D. rerio	334	88.3
C. elegans	131	85.5
D. melanogaster	143	93.0
A. thaliana	114	97.4
O. sativa	188	85.7
Total	**1479**	**85.1**

The fact that mouse miRNAs are predicted at lower rate probably reflects the larger number of hairpins registered for this species, many of which are not biochemically verified. Such discrepancies have been observed in other studies as well, although at a lesser extent (*e.g.*, [25]). The specificity over the mouse data is very high (84%) and remains surprisingly high in the two invertebrate species (~75%) (*data not shown*).

Comparison with other approaches

As described earlier, there are very few machine learning methods that do not require evolutionary information to predict miRNAs. To our knowledge, the only other probabilistic model is a motif finding method for mature miRNA region prediction [32]. An SVM-based approach has been proposed [25] that parses the *mfe* structure in "triplets": structural information about the pairing states of every three nucleotides, represented using dot-bracket notation. This method showed an accuracy of ~90% using the data available in the registry at the time. We used the same training and test sets used by the "triplet SVM" to train and test our model, HHMMiR, and we found it to perform better in almost all datasets (Table 5). The only exceptions are the mouse (but not rat) and *A. thaliana* (but not rice). Also, their model was able to predict all of the then five known miRNAs from Epstein-Barr virus, whereas HHMMiR predicted four. Overall, HHMMiR exhibits sensitivity of 93.2% and specificity of 89% in these datasets. If we limit the comparison of the two methods in one representative species from each taxon (*M. musculus*, *G. gallus*, *D. rerio*, *C. elegans*, *D. melanogaster*, *A. thaliana*, *Epstein Barr virus*) in order to minimize the statistical dependence of the data, the difference in the performance becomes statistically significant at the 5% level (p-value = 0.03, Wilcoxon paired test on the predicted number of genes).

Discussion

MiRNA genes constitute one of the most conserved mechanisms for gene regulation across all animal and plant species. The characteristics of the precursor miRNA stem-loops are well conserved in both vertebrate and invertebrate animals and fairly conserved between animals and plants. As seen in Table 1, plant hairpins tend to be generally longer than those in animals, while vertebrates have shorter precursors than invertebrates. Although, the "extension" and "pri-extension" regions may vary in length between animals and plants (much longer in plants), the lengths of the "miRNA" and "loop" regions are more similar. Thus, even across evolutionary time, the basic characteristics of miRNAs have not changed dramatically.

Table 5: Results for comparison between two precursor prediction methods.

Test Set	Total hairpins	Triplet SVM (%)	HHMMiR (%)
Positive Sets			
New human hairpins in registry at the time.	39	92.3	97.4
M. musculus	36	94.4	88.9
R. norvegicus	25	80.0	84.0
G. gallus	13	84.6	100
D. rerio	6	66.7	100
C. elegans	110	86.4	90.9
C. briggsae	73	95.9	95.9
D. melanogaster	71	91.6	95.8
D. pseudoobscura	71	90.1	98.6
A. thaliana	75	92.0	97.3
O. sativa	96	94.8	86.5
Epstein Barr virus	5	100	80.0
TOTAL	**620**	**91**	**93.2**
Negative Sets			
Folded genome hairpins from Chromosome 19	2444	89	88.6
Negative hairpin Set	1000	88.1	89.4
TOTAL	**3444**	**88.7**	**88.8**

The percentages represent the ratio of hairpins correctly predicted.

We designed a template for a typical precursor miRNA stem-loop and we built an HHMM based on it. HHMMiR was able to attain an average sensitivity of 84% and specificity of 88% on 10-fold cross validation of human data. We trained HHMMiR on human sequences and the resulting model was able to successfully identify a large percentage of not only mouse, but also invertebrate, plant and virus miRNAs (Table 4). This is an encouraging result showing that HHMMiR may be very useful in predicting miRNA genes across long evolutionary distances without the requirement for evolutionary conservation of sequences. This would be very beneficial for identification of miRNA hairpins in organisms that do not have closely related species sequenced, such as *Strongylocentrotus purpuratus* (sea urchin) and *Ornithorhynchus anatinus* (platypus) [33].

This is the first time a hierarchical probabilistic model has been used for classification and identification of miRNA hairpins. Probabilistic learning was previously applied by Nam *et al.* [32] for identifying the miRNA pattern/motif in hairpins. The advantage of the hierarchy used by our HHMMiR is that it parses each hairpin into four distinct regions and processes each of them separately. This represents better the biological role of each region, which is reflected in the distinct length distributions and neighbourhood base-pairing characteristic of that region. Furthermore, the underlying HHMM provides an intuitive modelling of these regions. We compared two modifications of the MLE and Baum-Welch algorithms for modelling the negative datasets, and we found them to perform similarly. Baum-Welch was selected for this study, since it does not require (random) labelling of the negative set.

The drawback of HHMMiR is that it depends on the *mfe* structure the *RNAfold* program returns [31]. In the future, we will test more folding algorithms or use the probability distribution of a number of top scoring folding energy structures returned by this package.

Conclusion

The success of our approach shows that the conservation of the miRNA mechanism may be at a much deeper level than expected. Further developments of the HHMMiR algorithm include the extension of the precursor template model (Figure 3) to be able to predict pri-miRNA genes with multiple stem-loops. Another extension would be to train a model to decode all HHMMiR predicted hairpins to identify the miRNA genes in them. Finally, it will be interesting to extend our method to include evolutionary information, which will allow us to assess the significance of conservation in predicting miRNA genes.

Methods

Data collection and processing

MiRNA dataset

MiRNA genes were obtained from the *microRNA registry*, version 10.1 (December 2007) [34], which contains 3265 miRNAs from animals and 870 from plants. For training HHMMiR, we used the residual 525 human hairpins, after filtering out precursor genes with multiple loops. Each gene was folded with the *RNAfold* program, which is part of the Vienna package [31], using the default parameters to obtain the secondary structure with minimum fold energy. The negative set consists of coding regions and random genomic segments from the human genome that were obtained using the UCSC genome browser [35]. These regions were folded and processed as described below.

Hairpin processing

Genomic sequences were folded in windows of 1 Kb, 500 nts and 300 nts with an overlap of 150 nts between consecutive windows. Nodes from the TeraGrid project [36] were used for this purpose. We tested the various window sizes on the relatively small *C. elegans* genome and discovered that 500 nts windows cover most known miRNA hairpins. Windows of 300 nts exhibited high degree of redundancy without adding more hairpins to those of the 500 nts windows, while 1 kb windows missed a higher percentage of known miRNAs (*data not shown*). For this study, we used hairpins extracted from windows of 500 nts. We were able to recover ~92% of the known miRNAs from *C. elegans* in this way. The remaining 8% may have been accounted for by existence of multiple loops or specificity of the parameters used. The hairpins were extracted from these folded windows using the following parameters: each hairpin has at least 10 base pairs, has a maximum length of 20 bases for the loop, and a minimum length of 50 nucleotides. The data flow of this process is presented in Figure 5.

After the hairpins are extracted, we process them to an input format representing the hairpin's secondary structure (Figure 5 and Figure 2) to be compatible with the HHMM shown in Figure 3. The labelling is done only for training data. For the purpose of labelling, the miRNA is first mapped to the folded hairpin (on either or both arms), and then the region representing the miRNA is labelled as the duplex miRNA (R) region. Our method does not consider the 3' overhangs generated during Dicer processing. The main bulge is labelled as the loop (L), whereas the remaining region between loop and miRNA is represented as the extension (E). The rest of the hairpin beyond the miRNA is labelled as pri-extension (P). A detailed description of these regions is given in the *Results* section.

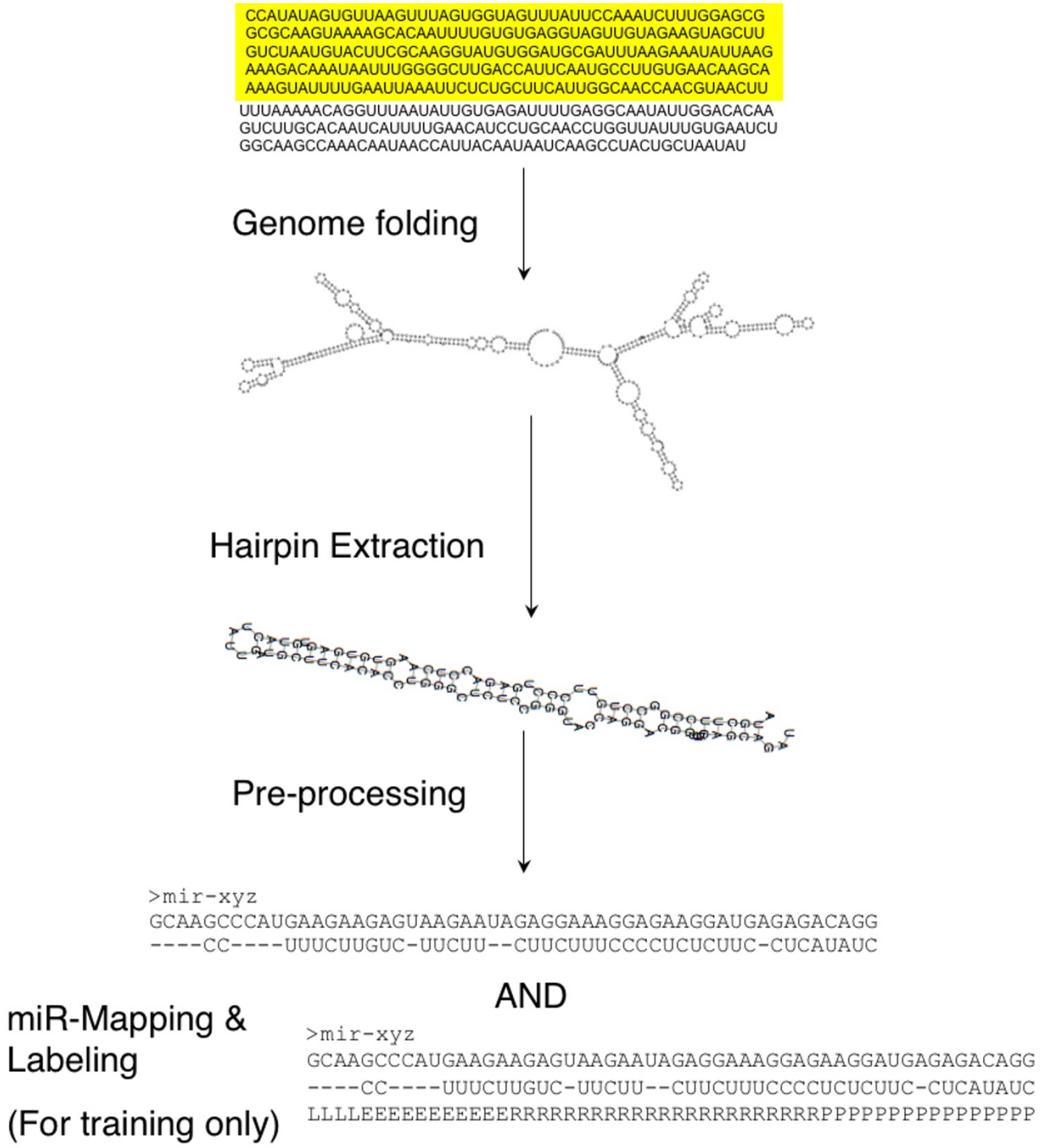

Figure 5
Data flow for hairpin extraction from the genome. The genome is first folded using windows of 500 nts with 150 nts overlap between consecutive windows. Hairpins are then extracted from the folded windows using the parameters described in the text. Hairpins are pre-processed into a suitable format for training/testing using the states shown in Figure 3 (L: Loop; E: Extension; R: miRNA; P: pri-miRNA extension). For the purpose of testing, the folded sequence is pre-processed into 2 lines of input representing the 2 stems of the hairpin. An example is given in Figure 2b.

Parameter estimation and testing

Parameter estimation

Two separate HHMM models are trained, one on positive data set (miRNAs and their corresponding hairpins) and the other on negative data set (hairpins, randomly chosen from the coding parts of the genome). The hairpins are pre-processed and labelled (if needed) before parameter estimation. Baum-Welch requires no labelling, but for MLE, we applied random labelling, as described above (Figure 2a).

The *alphabet* is denoted by $\Sigma = \{\sigma_i\}$ and the observed finite string is denoted by $\boldsymbol{O} = o_1 o_2 \ldots o_N$ such that $o_i \in \Sigma$. The i^{th} state at hierarchical level d is denoted as q_i^d (denoted as q^d in absence of ambiguity). The highest level of hierarchy (that of the root) is 1 while the lowest (that of the production states) is D (in our case, $D = 3$). The number of substates of each q_i^d ($d \in \{1, 2, \ldots D\text{-}1\}$ is $|q_i^d|$. The parameter set of an HHMM is denoted by:

$$\lambda = \left\{\lambda^{q^d}\right\}_{d\in\{1,\ldots,D\}} = \left\{\begin{array}{l} \left\{A\left(q^d\right)\right\}_{d\in\{1,\ldots,D\}}, \\ \left\{\prod\left(q^d\right)\right\}_{d\in\{1,\ldots,D\}}, \\ \left\{E\left(q^D\right)\right\} \end{array}\right\}$$

$A\left(q^d\right) = \left(a_{jk}^{q^d}\right) = P\left(q_k^{d+1} \middle| q_j^{d+1}\right)$ denoted by $\left\{A\left(q^d\right)\right\}_{d\in\{1,\ldots,D\}}$ is the state *transition matrix* of each internal substate, with $a_{jk}^{q^d} = P\left(q_j^{d+1} \middle| q_i^{d+1}\right)$ representing the probability that the j^{th} substate of q^d will transition to the k^{th} substate of q^d. Each internal state q^d has also an *initial distribution vector* $\prod\left(q^d\right) = \left\{\pi\left(q_j^{d+1} \middle| q^d\right)\right\} = \left\{P\left(q_j^{d+1} \middle| q^d\right)\right\}$ denoted by $\left\{\prod\left(q^d\right)\right\}_{d\in\{1,\ldots,D\}}$, where $\pi\left(q_j^{d+1} \middle| q^d\right)$ is the probability that q^d will make a vertical transition to its j^{th} substate at level d+1, thus, activating it. The production states q^D will have *emission probability vector* or the *output distribution vector*

$$E\left(q^D, q^{D-1}\right) = \left\{e\left(\sigma_l \middle| q^D, q^{D-1}\right)\right\} = \left\{P\left(\sigma_l \middle| q^D, q^{D-1}\right)\right\}$$

denoted by $\{E(q^D)\}$ where $e(\sigma_l|q^D, q^{D\text{-}1})$ is the probability that production state q^D will emit symbol $\sigma_l \in \Sigma$.

Now we will define the various probabilities that are required to be calculated for parameter estimation.

(*i*) $\alpha\left(t, t+k, q_i^{d+1}, q^d\right) = P(o_t \cdots o_{t+k}, q_i^{d+1}$ finished at $o_{t+k} \middle| q^d$ started at o_t) is the *forward probability* of emitting the substring $o_t \ldots o_{t+k}$ of the observation sequence by the parent state q^d such that it was entered at o_t and the subsequence ended at substate q_i^{d+1} and thus, it was the last state activated.

(*ii*) $\chi\left(t, q_i^{d+1}, q^d\right)$ is the probability of making a vertical transition from parent q^d to q_i^{d+1} just before the emission of o_t.

(*iii*) $\xi\left(t, q_i^{d+1}, q_j^{d+1}, q^d\right) = P(o_1 \cdots o_t, q_i^{d+1} \to q_j^{d+1}, o_{t+1} \cdots o_N \mid \lambda)$ is the probability of making a horizontal transition from

Table 6: Measures for accuracy calculation.

Measure	Calculation
Sensitivity (Sn)	$Sn = TP/(TP + FN)$
Specificity (Sp)	$Sp = TN/(TN + FP)$
False Discovery Rate (FDR)	$FDR = FP/(TP + FP)$
Matthew's Correlation Coefficient (MCC)	$MCC = (TP \cdot TN - FP \cdot FN)/\left(\sqrt{(TP + FP)\cdot(TP + FN)\cdot(TN + FP)\cdot(TN + FN)}\right)$

TP: *True Positives*; TN: *True Negatives*; FP: *False Positives*; FN: *False Negatives*.

q_i^{d+1} to q_j^{d+1} where both are substates of q^d after the emission of o_t and before the emission of o_{t+1}.

(*iv*) $\gamma_{in}\left(t, q_i^{d+1}, q^d\right) = \sum_{k=1}^{|q^d|} \xi\left(t-1, q_k^{d+1}, q_i^{d+1}, q^d\right)$ is the probability of performing a horizontal transition to q_i^{d+1} which is substate of q^d before o_t is emitted. Further details on the algorithms are given in [27] and in Additional file 2.

The parameters are estimated as follows:

$$\hat{\pi}\left(q_i^2 \middle| q^1\right) = \chi\left(t, q_i^2, q^1\right)$$

$$\hat{\pi}\left(q_i^{d+1} \middle| q^d\right) = \frac{\sum_{t=1}^{T} \chi\left(t, q_i^{d+1}, q^d\right)}{\sum_{i=1}^{|q^d|} \sum_{t=1}^{T} \chi\left(t, q_i^{d+1}, q^d\right)} \quad (1 < d < D-1)$$

$$\hat{a}_{jk}^{q^d} = \frac{\sum_{t=1}^{T} \xi\left(t, q_i^{d+1}, q_j^{d+1}, q^d\right)}{\sum_{k=1}^{|q^d|} \sum_{t=1}^{T} \xi\left(t, q_i^{d+1}, q_k^{d+1}, q^d\right)}$$

$$\hat{e}\left(\sigma_l \middle| q^D, q^{D-1}\right) = \Big(\sum_{o_t=\sigma_l} \chi\left(t, q_i^D, q^{D-1}\right) + \sum_{t>1, o_t=\sigma_l} \gamma_{in}\left(t, q_i^D, q^{D-1}\right)\Big) \Big/ \Big(\sum_{t=1}^{T} \chi\left(t, q_i^D, q^{D-1}\right) + \sum_{t=2}^{T} \gamma_{in}\left(t, q_i^D, q^{D-1}\right)\Big)$$

Testing

As described above, classification of test hairpins depends on the ratio of the log-likelihoods generated by the positive and negative models. A threshold was decided for this ratio using the ROC curves shown in Figure 4. For each hairpin, the probability that a certain model emitted the hairpin is given by:

$$P\left(O \middle| \lambda\right) = \sum_{i=1}^{|q^1|} \alpha\left(1, T, q_i^2, q^1\right)$$

Measures of accuracy

The different terms and measures used to calculate the efficiency of HHMMiR are listed in the Table 6.

List of abbreviations used

HHMM: hierarchical hidden Markov model; mfe: minimum fold energy; miRNA: microRNA; MLE: maximum likelihood estimate.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PVB and SK designed the study, analyzed the results and wrote the paper. SK implemented the HHMM. VH supervised the data analysis and contributed to the writing of the paper.

Additional material

Additional File 1

This file contains the results of summarization of the microRNA registry *(version 10.1, December 2007) [34] hairpin characteristics for each species.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S1-S35-S1.xls]

Additional File 2

This file contains a more detailed description of the algorithms used for parameter estimation and classification using HHMMs.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S1-S35-S2.pdf]

Acknowledgements

The authors would like to thank Paul Samollow, Chakra Chennubhotla, Eleanor Feingold and an anonymous reviewer for helpful suggestions. PVB was supported by NIH grant 1R01LM009657-01. This research was supported in part by the National Science Foundation through TeraGrid resources provided by Pittsburgh Supercomputing Center. Supplementary material can be found at the journal's web site and at our web site [37].

This article has been published as part of *BMC Bioinformatics* Volume 10 Supplement 1, 2009: Proceedings of The Seventh Asia Pacific Bioinformatics Conference (APBC) 2009. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/10?issue=S1

References

1. Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN: **MicroRNA genes are transcribed by RNA polymerase II.** *Embo J* 2004, **23(20):**4051-4060.
2. Cai X, Hagedorn CH, Cullen BR: **Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs.** *Rna* 2004, **10(12):**1957-1966.
3. Yi R, Qin Y, Macara IG, Cullen BR: **Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs.** *Genes Dev* 2003, **17(24):**3011-3016.
4. Chendrimada TP, Gregory RI, Kumaraswamy E, Norman J, Cooch N, Nishikura K, Shiekhattar R: **TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing.** *Nature* 2005, **436(7051):**740-744.

5. Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, *et al.*: **The nuclear RNase III Drosha initiates microRNA processing.** *Nature* 2003, **425(6956):**415-419.
6. Schwarz DS, Hutvagner G, Du T, Xu Z, Aronin N, Zamore PD: **Asymmetry in the assembly of the RNAi enzyme complex.** *Cell* 2003, **115(2):**199-208.
7. Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W, Tuschl T: **Identification of tissue-specific microRNAs from mouse.** *Curr Biol* 2002, **12(9):**735-739.
8. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.** *Cell* 2005, **120(1):**15-20.
9. Millar AA, Waterhouse PM: **Plant and animal microRNAs: similarities and differences.** *Functional & integrative genomics* 2005, **5(3):**129-135.
10. Sarnow P, Jopling CL, Norman KL, Schutz S, Wehner KA: **MicroRNAs: expression, avoidance and subversion by vertebrate viruses.** *Nature reviews* 2006, **4(9):**651-659.
11. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.** *Cell* 1993, **75(5):**843-854.
12. Wightman B, Ha I, Ruvkun G: **Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans.** *Cell* 1993, **75(5):**855-862.
13. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G: **The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans.** *Nature* 2000, **403(6772):**901-906.
14. Berezikov E, Cuppen E, Plasterk RH: **Approaches to microRNA discovery.** *Nat Genet* 2006, **38(Suppl):**S2-7.
15. Yang W, Chendrimada TP, Wang Q, Higuchi M, Seeburg PH, Shiekhattar R, Nishikura K: **Modulation of microRNA processing and expression through RNA editing by ADAR deaminases.** *Nature structural & molecular biology* 2006, **13(1):**13-21.
16. Yang Z, Ebright YW, Yu B, Chen X: **HEN1 recognizes 21–24 nt small RNA duplexes and deposits a methyl group onto the 2' OH of the 3' terminal nucleotide.** *Nucleic Acids Res* 2006, **34(2):**667-675.
17. Ohler U, Yekta S, Lim LP, Bartel DP, Burge CB: **Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification.** *Rna* 2004, **10(9):**1309-1322.
18. Grad Y, Aach J, Hayes GD, Reinhart BJ, Church GM, Ruvkun G, Kim J: **Computational and experimental identification of C. elegans microRNAs.** *Mol Cell* 2003, **11(5):**1253-1263.
19. Lai EC, Tomancak P, Williams RW, Rubin GM: **Computational identification of Drosophila microRNA genes.** *Genome Biol* 2003, **4(7):**R42.
20. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP: **The microRNAs of Caenorhabditis elegans.** *Genes Dev* 2003, **17(8):**991-1008.
21. Legendre M, Lambert A, Gautheret D: **Profile-based detection of microRNA precursors in animal genomes.** *Bioinformatics* 2005, **21(7):**841-845.
22. Sewer A, Paul N, Landgraf P, Aravin A, Pfeffer S, Brownstein MJ, Tuschl T, van Nimwegen E, Zavolan M: **Identification of clustered microRNAs using an ab initio prediction method.** *BMC Bioinformatics* 2005, **6:**267.
23. Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E, *et al.*: **Identification of hundreds of conserved and nonconserved human microRNAs.** *Nat Genet* 2005, **37(7):**766-770.
24. Pfeffer S, Sewer A, Lagos-Quintana M, Sheridan R, Sander C, Grasser FA, van Dyk LF, Ho CK, Shuman S, Chien M, *et al.*: **Identification of microRNAs of the herpesvirus family.** *Nat Methods* 2005, **2(4):**269-276.
25. Xue C, Li F, He T, Liu GP, Li Y, Zhang X: **Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine.** *BMC Bioinformatics* 2005, **6:**310.
26. Terai G, Komori T, Asai K, Kin T: **miRRim: a novel system to find conserved miRNAs with high sensitivity and specificity.** *Rna* 2007, **13(12):**2081-2090.
27. Fine S, Singer Y, Tishby N: **The Hierarchical Hidden Markov Model: Analysis and Applications.** *Machine Learning* 1998, **32:**41-62.
28. Hu M, Ingram C, Sirski M, Pal C, Swamy S, Patten C: **A Hierarchical HMM Implementation for Vertebrate Gene Splice Site Prediction.** In *Technical Report Department of Computer Science, University of Waterloo*; 2000.
29. Lin T, Ray P, Sandve GK, Uguroglu S, Xing EP: **BayCis: A Bayesian Hierarchical HMM for Cis-Regulatory Module Decoding in Metazoan Genomes.** In *Research in Computational Molecular Biology (RECOMB), 12th Annual International Conference: 2008 Singapore: Springer*; 2008:66-81.
30. Saetrom P, Snove O, Nedland M, Grunfeld TB, Lin Y, Bass MB, Canon JR: **Conserved microRNA characteristics in mammals.** *Oligonucleotides* 2006, **16(2):**115-144.
31. Hofacker IL: **Vienna RNA secondary structure server.** *Nucleic Acids Res* 2003, **31(13):**3429-3431.
32. Nam JW, Shin KR, Han J, Lee Y, Kim VN, Zhang BT: **Human microRNA prediction through a probabilistic co-learning model of sequence and structure.** *Nucleic Acids Res* 2005, **33(11):**3570-3581.
33. Samollow PB: **The opossum genome: insights and opportunities from an alternative mammal.** *Genome Res* 2008, **18(8):**1199-1215.
34. Griffiths-Jones S: **miRBase: the microRNA sequence database.** *Methods Mol Biol* 2006, **342:**129-138.
35. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12(6):**996-1006.
36. Catlett C, Allcock WE, Andrews P, Aydt R, Bair R, Balac N, Banister B, Barker T, Bartelt M, Beckman P, *et al.*: **TeraGrid: Analysis of Organization, System Architecture, and Middleware Enabling New Types of Applications.** In *High Performance Computing and Grids in Action Volume 16*. Edited by: *Grandinetti L. Amsterdam: IOS Press*; 2008.
37. **Benos laboratory web server** [http://www.benoslab.pitt.edu/]