Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

PATRICK MAY, MARKUS BAUER, CHRISTIAN KÖBERLE AND
GUNNAR W. KLAU

A computational approach to microRNA detection

ZIB-Report 06-44 (November 2006)

A computational approach to microRNA detection

Patrick May[1,2], Markus Bauer[2,3], Christian Köberle[4], and Gunnar W. Klau[5,6]

[1] *Computer Science Research, Dept. Computer Science, Zuse Institute Berlin,*

[2] *Algorithmic Bioinformatics, Dept. Computer Science, Free University, Berlin, Germany,*

[3] *International Max Planck Research School, Berlin, Germany,*

[4] *Max Planck Institute for Infection Biology, Berlin, Germany,*

[5] *Mathematics in Life Sciences, Dept. Mathematics, Free University, Berlin, Germany,*

[6] *DFG Research Center* Matheon*, Berlin, Germany*

February 20, 2007

Abstract

During the last few years more and more functionalities of RNA have been discovered that were previously thought of being carried out by proteins alone. One of the most striking discoveries was the detection of microRNAs, a class of noncoding RNAs that play an important role in post-transcriptional gene regulation. Large-scale analyses are needed for the still increasingly growing amount of sequence data derived from new experimental technologies. In this paper we present a framework for the detection of the distinctive precursor structure of microRNAS that is based on the well-known Smith-Waterman algorithm and various filtering steps. We conducted experiments on real genomic data and we found several new putative hits for microRNA precursor structures.

Contents

1 Introduction

One of the most exciting recent discoveries in molecular genetics is the important function of noncoding RNAs (ncRNA) in various catalytic and regulatory processes (see *e.g.* [1, 2]). Many experimental studies have shown that large fractions of the transcriptome are consisting of ncRNAs [3–6]. Additionally, there are several bioinformatics surveys providing evidence for a large amount of ncRNAs in various species [7–9]. Due to the ongoing sequencing projects and the large amount of data coming from molecular biology techniques like tiling arrays and cDNA sequencing one of the major challenges in bioinformatics and computational biology will be large-scale analysis and predictions of ncRNAs. Such high-throughput, genome-wide searches require fast and efficient detection algorithms as well as high-performance technologies.

MicroRNAs form a conserved, endogenous 21-23 nucleotide class of ncRNAs. They regulate protein-coding gene expression in plants and animals via the RNA silencing machinery (reviewed in [10,11]). Recently, microRNAs have been also discovered in viral genomes indicating that viruses have evolved to regulate both host and viral genes [12]. MicroRNAs are derived from larger, approximately 70 nt long, precursors that form imperfect stem-loop structures. In the cytoplasm the precursor is cleaved by the enzyme Dicer to excise the mature microRNA, which is assembled into the RNA silencing machinery. By binding to a complementary target in the mRNA, the mature microRNA inhibits translation or facilitates cleavage of the mRNA [13].

Computational methods that search for ncRNAs can be divided in two classes: approaches that try to detect ncRNAs *de novo* (like `RNAz` [14] or `QRNA` [15]), and programs that search for homologous structures in genomic sequences, given the structure of a known ncRNA family. Programs for the latter task include, *e.g.*, `FastR` [16] or `RSEARCH` [17]. `FastR` takes as its input a so called *filter mask* that basically specifies distance constraints between the various stems. It then searches for exact matching stems of a certain length (the default value is 3) that satisfy these constraints: if all stems are found, the specific region is evaluated again by a more costly alignment algorithm. Since only exact-matching stems are searched, this task can efficiently be done using hash tables, resulting in a very fast filter stage.

`RSEARCH` on the other hand works with *covariance models*: Covariance models are stochastic context-free grammars that capture both the sequence and the structure of a given training set. Once the parameters of the models are evaluated, it can be used to align the query with a target sequence. `RSEARCH` provides high specificity and sensitivity at the expense of high computational demands.

Whereas the approaches mentioned above deal with general RNA structures, there are several other approaches that especially focus on the detection of putative microRNA precursor structures. Each one of these, however, relies on additional information in order to increase specificity and sensitivity: `RNAmicro` [18] depends on a given multiple alignment—which is not always available, e.g., in case of viruses—to detect conserved microRNA precursor structures using essentially the same algorithm as described in [14]. The tools `harvester` [19] and `miralign` [20] search for homologous structures given a set of known microRNA precursors. They use sequence and structure features to classify hairpin structures. The `miR-abela` software [21] computes statistical features of hairpin loops and passes these values to a *support vector machine* as its input. Hence, `miR-abela` relies partly on known microRNAs to find homologous novel candidates. Furthermore, in order to limit the number of possible hits, the search is restricted to the neighborhood of confirmed microRNA hits (the authors reason that microRNAs tend to appear in clusters and the search for novel candidates should therefore be conducted in vicinity to known microRNAs; this assumption is backed up by the computational results conducted in [18]).

In this paper we present a hierarchical framework for the detection of the distinctive precursor structure of microRNAs using the Smith-Waterman algorithm. The pipeline consists of several filtering steps, the most important of which employ the Smith-Waterman dynamic programming algorithm for local sequence alignment. We also show how to use the pipeline in a comparative approach that employs the strong evolutionary conservation of microRNA structures but does not depend on multiple alignments. We conducted experiments on real genomic data and we analysed all known metazoan and virus microRNAs from the mirBase in the context of our proposed method to show that we are able to predict putative microRNA precursor candidates.

2 Results and Discussion

In this section we present the filter steps of our method and the results we obtained on the microRNA sequence data base miRBase [22].

2.1 Scanning a single genomic sequence: an overview

The main idea of our hierarchical filter approach is to design a computational method that searches for basic features of microRNA precursors by applying various filter stages of increasing efficiency at the expense of increasing running time. Hereby, we use several general structural observations about microRNA precursor structures. We define a microRNA precursor candidate according to the structural criteria that were defined by Ambros *et al.* [23]. The microRNA precursor structure must have a characteristic secondary structure forming a hairpin or fold-back with a stem that can contain small internal loops or bulges, and which is important for enzymatic recognition [24, 25] or transportation out of the nucleus [26]. Figure 1 shows some example microRNA precursor candidates predicted by our method. The mature microRNA has to be part of one arm within the hairpin structure. Bonnet *et al.* [27] have found that microRNA precursors, unlike other ncRNAs like transfer RNAs or ribosomal RNAs, have significantly lower minimum free folding energies (MFEs) than random sequences. Additionally, the predicted microRNA precursor structure should be robust with respect to its surrounding genomic context, because the processed intermediates of the microRNA pathway can adopt various lengths [21].

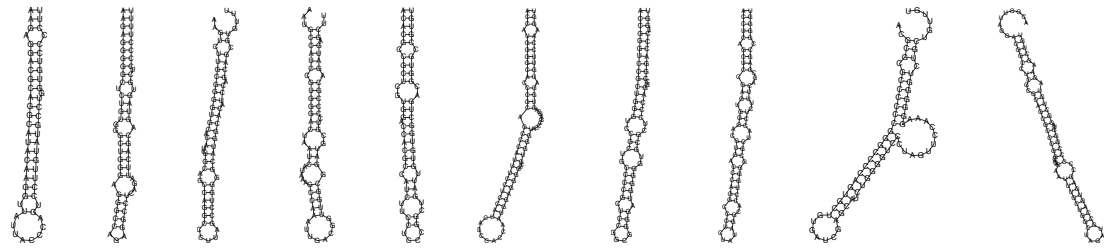


Figure 1: Typical microRNA precursor structures detected by our computational pipeline. The picture (generated with RNAfold [28]) shows the first ten predicted precursor candidates of the Epstein-Barr virus genome, including known microRNA precursors (a) `ebv-mir-BART17` (miRBase [22] accession id `MI0004990`), (n) `ebv-mir-BART13` (id `MI0003735`), and (o) `ebv-mir-BART6` (id `MI0003728`).

Our method has one or more genomic sequences as input. Then, for each sequence, we apply the following filtering steps, also illustrated in Fig. 2:

1. Since the secondary structure of a microRNA precursor is a simple stem-loop structure, we are first searching genomic regions that are likely to exhibit the hairpin structure. To this end, we are using *local alignment* to approximate the folding process (see Methods for details). If the local alignment score is beyond a certain threshold, we include the corresponding subsequence in a list of putative microRNA precursor candidates.

2. A window resulting in a good local alignment score is likely to have neighboring windows with similarly good scores. We therefore keep only the best-scoring of overlapping precursor candidates.

3. We have observed that some of the remaining candidates occur more than once on the genomic sequence. We agglomerate these candidates and leave it up to the user whether to keep only a representative candidate of the repeated hits or not.

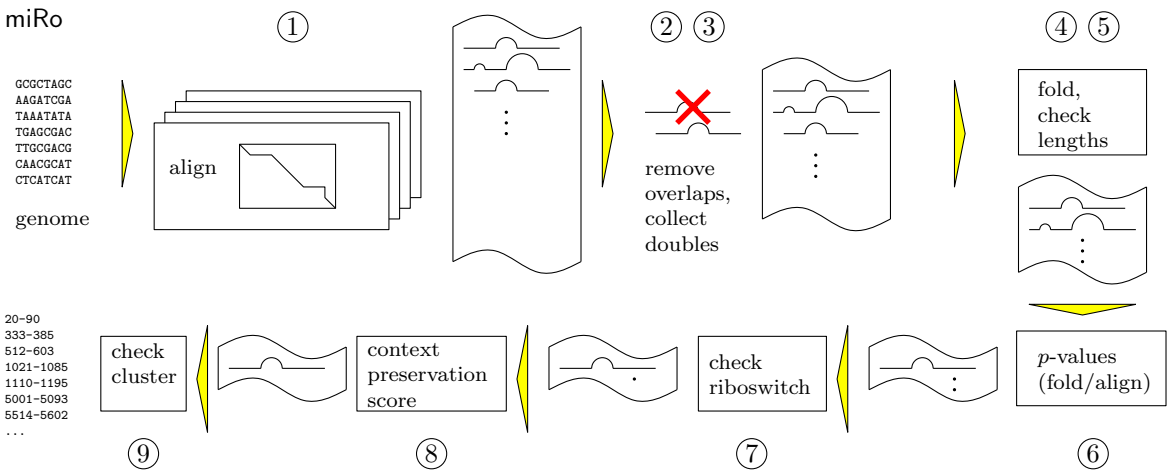


Figure 2: Overview of the filtering approach on a single genomic sequence.

4. We now fold each candidate on the list (currently with `RNAfold` [28]) and keep only those candidates whose predicted structure corresponds to a stem-hairpin structure. Therefore, we inspect the predicted structure and keep a candidate only if the longest stem-hairpin structure satisfies the stem-loop criteria by parsing the bracket notation. Additionally, we discard a candidate if its predicted MFE is above a certain threshold.

5. After folding we check whether the longest helical arm is long enough to accommodate a microRNA. If not, we remove the candidate.

6. We compute two p-values for the remaining microRNA precursor candidates randomizing the sequence using dinucleotide shuffling (see Methods for details). If these p-values exceed certain thresholds, we discard the corresponding candidate.

7. After sorting the microRNA precursor candidates by their genomic positions we remove those candidates that are overlapping. The removed hits possibly point to candidates for other types of ncRNAs, like for instance riboswitches.

8. Now, we are looking if the microRNA precursor hairpin structure is conserved within the sequential and structural context of the genomic sequence. Therefore, we are adding at both sides of the candidate the original genomic sequence of length *context*, fold the elongated candidate, and calculate the *context-preservation* score (see Methods for details). If the original candidate sequence is not folding into a hairpin structure anymore, we disregard the candidate.

9. Since microRNAs tend to reside in clusters [29] we optionally remove all candidates that have no neighbor within a *maxdist* radius, where *maxdist* is the maximal allowed sequence separation between two microRNAs within the same gene cluster. Clusters containing only single microRNA candidates are optionally removed.

2.2 Local alignment vs. folding

To evaluate if our local alignment method is able to detect stem-loop candidates in various genomic sequences, we first analysed all metazoan and virus microRNA precursor sequences that were annotated in the miRBase sequence database, release 9.0. Table 1 gives the basic statistical analysis from totally 3498 sequences from 40 species. For every microRNA precursor we first determine the MFE, the MFE p-value, its sequence length, and the number of paired nucleotides. All calculations were done with RNAfold [28] standard energy

Table 1: Basic statistic values for the all metazoan and virus sequences used in this paper. Statistical analysis for 3498 microRNAs. Shown are the mean values per species for the original microRNA precursors and the best candidates we have found: the number of microRNA precursor (No) number of base pairs (BP, BPc), length (L, Lc), minimum free energy (MFE, MFEc), p-values for MFE (MFE-Pv, MFEc-Pv), alignment score (Sc), p-value for alignment score (Sc-Pv).

Species	No.	BP	BPc	L	Lc	MFE	MFEc	MFE-Pv	MFEc-Pv	Sc	Sc-Pv
Arthropoda	235	31	30	90	83	-36	-34	0.0042	0.0138	45	0.0178
Nematoda	193	34	30	99	84	-38	-33	0.0158	0.0520	46	0.0136
Platyhelminthes	63	33	31	99	84	-42	-38	0.0010	0.0025	45	0.0163
Vertebrata	2325	31	30	87	80	-38	-36	0.0180	0.0283	46	0.0334
Pisces	600	30	29	85	78	-36	-34	0.0061	0.0131	45	0.0326
Viruses	82	28	27	81	76	-36	-35	0.0087	0.0096	45	0.0262
Total	3498	32	30	88	80	-37	-35	0.0166	0.0326	46	0.0294

parameters. The results from our study are very simular to those from Bonnet *et al.* [27]. Afterwards we calculated for this precursor sequence the best candidate according our proposed local alignment strategy. For the best candidate we determined again MFE, MFE p-value, sequence length, number of paired nucleotides, and additionally the p-value for the local alignment score. Although our best candidates for the original microRNA precursors were significantly shorter, the MFEs for the original sequence and the best candidate sequence are very close. The p-value for the original MFE is only slightly better than the p-value of the alignment score which indicates that the simulation of the folding process of the simple stem-loop structure by our alignment method is a reasonable method. Additionally, the distributions of p-values (Figure 3) show that using the alignment shuffling as a first filter will reduce the number of false-positives significantly so that we can afterwards proceed with calculating the MFE p-value, a more accurate, but also more time-consuming filter step.

The actual p-values resulting from the original MFE and alignment scores are always relatively close to each other, yielding the question whether the two p-values do not statistically differ. Since we do not know anything about the actual distribution of our scores, we chose the nonparametric *Wilcoxon Signed-Rank* test to determine whether the two sets of scores are statistically the same. Taking the MFE and alignment p-values of all 3498 confirmed microRNAs from Table 1 the Wilcoxon Signed-Rank test gives a p-value of $\approx 10^{-16}$ showing that the two scores are statistically not different from each other. This again implies that using the alignment p-value alone (which is much faster to compute) can be sufficient for the evaluation of putative microRNA candidates.

Another advantage of our local alignment strategy is the running time. In comparison to our local alignment strategy folding the subsequence of a greater genomic sequence induced by a window has two major drawbacks: First, folding algorithms are computationally quite expensive—the most popular algorithms take $O(n^3)$ time, where n is the length of the folded sequence. Experiments on AMD 2.2 GHz Opteron CPUs showed that processing a sequence of about 130.000 nucleotides with a sliding window size of 90 nt and a step size of one nucleotide took about 28 minutes when folding the window with RNAfold [28] but only about 1.5 minutes using our local alignment approach based on the Smith-Waterman algorithm method as described in the Methods section. Second, the MFE of a sequence grows with the sequence length. Algorithms based solely on folding must therefore choose the window size very carefully as shorter structures within the window tend to be overlooked.

2.3 Context folding

To support the idea that microRNA precursor are robust with respect to their genomic context, we retrieved the available mammalian microRNAs from the miRBase (see [22]), yielding 491 confirmed hits (we were unable to find the remaining confirmed microRNAs in the genomes specified). We then added 180 nucleotides

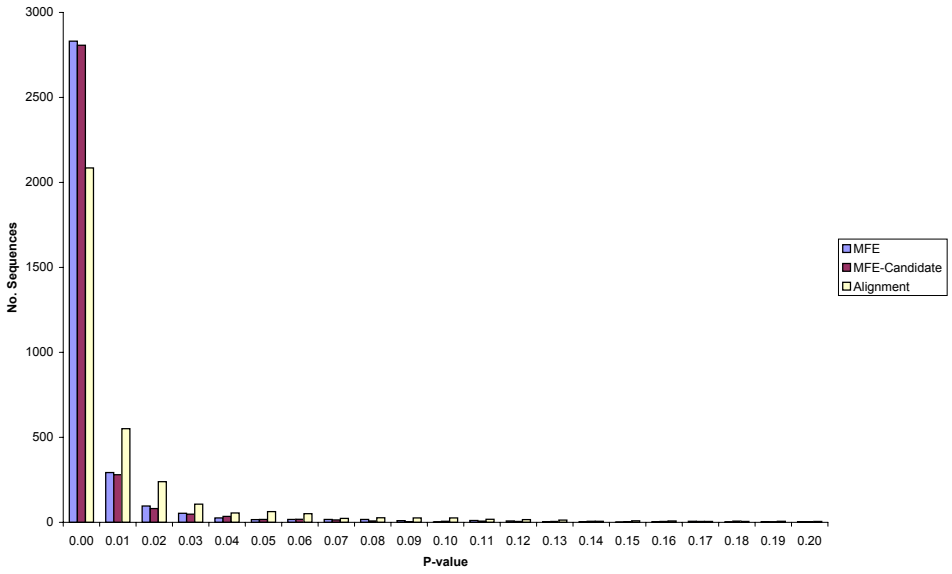


Figure 3: Distribution of p-values for original MFE, best candidate MFE, and best candidate alignment score

to both sides of the original microRNA from its genomic context, and computed the *context-preservation score*, which should reflect how robust a hairpin is relative to its surrounding genomic context (see Methods for details). We applied the same procedure to 746 of our putative precursor structures predicted for the Epstein-Barr virus genome. Figure 4 shows the distribution of the context-preservation scores of confirmed and putative microRNAs, respectively. Furthermore, we were interested if the size of the additional genomic context had an influence on the score distribution. The plot on the right-hand side of Figure 4 shows the different distributions of the scores depending on the number of added nucleotides (either 50, 120, 180, or 300). Surprisingly, the values are almost identically distributed for different numbers of additional nucleotides. This in turn implies that we can restrict ourselves to a smaller genomic context for the computation of the context-preservation score which leads to significantly reduced running times (remember that folding has a complexity of $\mathcal{O}(n^3)$): as an illustration, computing the context-preservation score for 100 putative microRNAs with a genomic context of 120 nucleotides to both sides takes 80 seconds, whereas it already grows to 576 seconds in the case of 300 nucleotides.

2.4 Filter performance

Our goal was to assess the "power" of each single filter step. To this end, we shuffled the genome of the Epstein-Barr virus such that the dinucleotide frequency was preserved, and planted the known 23 microRNAs into the shuffled sequence afterwards. We then searched the shuffled sequence (containing the original unshuffled 23 Epstein-Barr virus microRNAs) and the original Epstein-Barr virus genome for candidate half stems and applied the filter stages separately. Scanning the sequences for candidate hairpin structures (i.e.

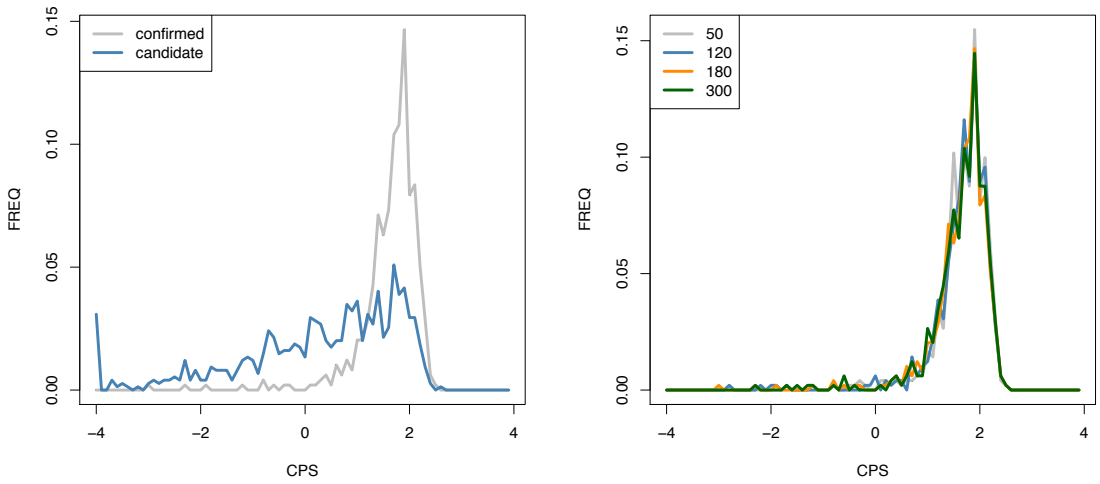


Figure 4: In the left-hand figure the blue and grey line show the distribution of context-preservation scores of 491 confirmed microRNAs and microRNA candidates as detected by our algorithm. Although the two distributions do overlap, one can clearly see that the distribution of the confirmed microRNAs has a peak at around 2, whereas the candidates are distributed over the entire spectrum. The right-hand side shows the CPS distribution for different numbers of added nucleotides.

without applying any filter stages) resulted in a set of 927 candidates for the shuffled sequence and 1071 candidates for the original genome after the local alignment procedure. Table 2 gives an overview on the performance of the filter stages described above. For both sequences the MFE-p-value was the most efficient filtering step, but also the most time-consuming, followed by the alignment p-value calculation.

The results described in Table 2 suggest the order of filter steps as illustrated in Fig. 2, which is also the default order in our implementation. Note that this can be easily changed and filter steps may be optionally switched on or off. Beyond, our implementation is open to additional filter steps, e.g., GC content or nucleotide frequencies which are also used in similar methods [21].

Observe that we did not apply the context-preservation filter, because this filter stage depends on the genomic context of the microRNA (which is random in this case), and we did not do the cluster filter stage (because the known microRNA were randomly distributed in the shuffled sequence).

2.5 Searching microRNA precursor structures

To evaluate the ability of miRo to search for microRNA precursor in single genomes, we first tested it on the Epstein-Barr virus genome. The miRo tool detected all 23 currently annotated microRNAs from the miRBase database [22] using standard parameters. Our method reported 49 new putative miRNA precursor candidates. The precursors of the 23 known microRNAs can be grouped into four clusters when using a maximum allowed distance of 1500 between two microRNA precursor sequences to define a cluster:

1. position 41471 – 43030 containing 3 microRNAs (*cluster*1),
2. position 139076 – 140107 containing 8 microRNAs (*cluster*2),
3. position 145932 – 148815 containing 11 microRNAs (*cluster*3), and
4. position 152745 – 152806 containing 1 microRNA (*cluster*4).

Table 2: Performance of the different miRo filter stages on the initial sets of putative microRNA precursor structures from the Epstein-Barr virus (*ebv*) and the the shuffled Epstein-Barr virus (*ebvs*). The numbers in brackets correspond to the filter stages as denoted in Figure 2.

	ebv			**ebvs**		
Filter stage	**No. hits**	**Missed hits**	**Time (sec)**	**No. hits**	**Missed hits**	**Time (sec)**
Low alignment score (2)	1061	0	59	911	0	53
Repetitive microRNAs (3)	743	0	56	927	0	54
Bad folded (4)	774	0	58	640	0	59
Low MFE score (4)	825	0	60	529	0	53
Short half stem (5)	773	0	58	659	0	70
Alignment p-value (6)	649	0	172	637	5	136
MFE p-value (6)	440	1	669	331	1	587

With the miRo algorithm we could annotate *cluster*1 and *cluster*2 with each one additional microRNA precursor candidate. The clusters *cluster*3 and *cluster*4 could be combined into one single gene cluster by adding eight new candidates. Moreover, we found seven new clusters of potential microRNAs precursors: four clusters containing 3 candidates, one cluster containing 5 candidates, one cluster containing 6 candidates, and one cluster containing 9 candidates.

2.6 Comparative application

Applying miRo on a single genomic sequence usually leads to a large list of candidates, which contains still many false positives. This is due to two major reasons:

- It has been verified that large parts of a genomic sequence can fold into a microRNA-like structure. For instance, there are up to eleven million hairpin structures in the entire human genome [30]. Many of these hits are still contained in the final candidate list.
- The list still contains stem-hairpin structures that may be part of a larger non-coding RNA.

We therefore propose a comparative approach (see Fig. 5 and Methods section) to remove most of these false positive hits. The key idea is to apply miRo to different genomes and then compare the candidates in the output lists against each other. For the local alignment phase (Fig. 5) we used an alignment scoring based on matching nucleotides: +1 for identity, -2 for non matching nucleotides, and -2 for gaps.

We wish to emphasize that our comparative approach, in contrast to other methods, does not rely on good multiple alignments. Often—*e.g.*, in the case of viral genomes—alignments just do not exist, in many other cases, the quality of the multiple alignments is bad due to the difficulty of computing reliable structural alignments.

Since microRNAs are often highly conserved across closely related species, we tried to search highly conserved half stems of microRNA precursor candidates on five herpesvirus genomes (see Table 3).

Table 3: Genomes used in the experiments. For every genome the number of nucleotides, the accession number and a short description is listed.

Short name	**Size**	**Genbank Id**	**Genome Description**
Epstein-Barr	171823	86261677	Human herpesvirus 4 wild type
Herpes1	152261	9629378	Human herpesvirus 1
Herpes2	230287	28373214	Herpesvirus 5 (laboratory strain AD169)
Herpes3	154746	9629267	Herpesvirus 2
Herpes4	159321	9628290	Herpesvirus 6

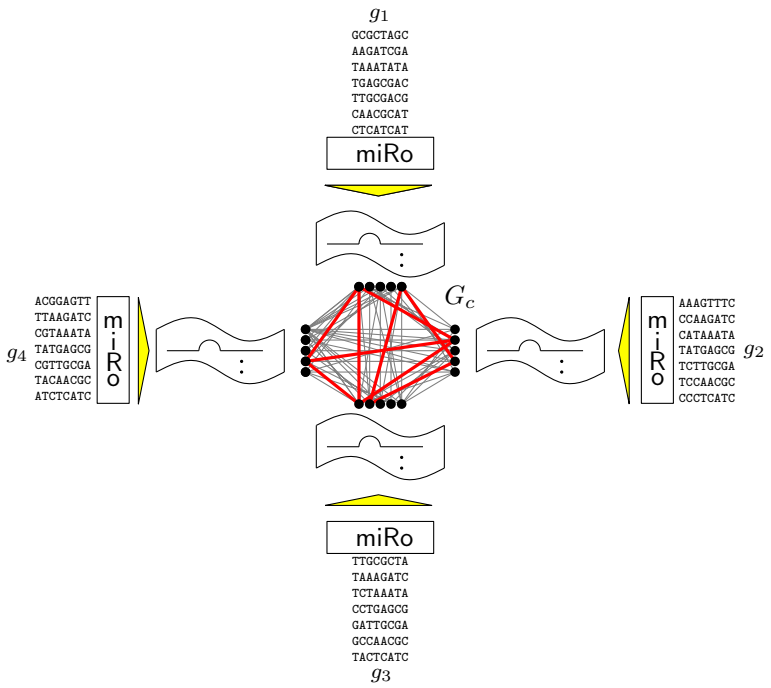


Figure 5: Overview of the comparative approach for four genomes $g_1, \ldots, g_4$. In the example, the conservation graph G_c contains two maximal cliques of size four and three, respectively.

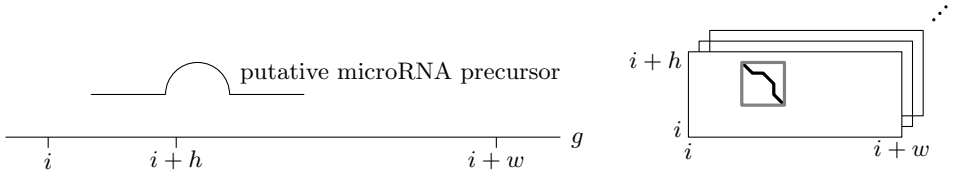
We found 18 conserved microRNA precursor candidates between *Herpes 1* and *Herpes 3* genomes and one conserved microRNA precursor candidate between *Herpes 2* and *Herpes 3* genomes. Since no microRNAs are known for this genomes it would be very interesting to cooperate with laboratory groups to look if these candidates could be experimentally validated.

3 Methods

3.1 Searching microRNA candidates using local alignment

Similar to other approaches, we slide a window of length w positionwise along the given genomic sequence g and check whether we detect a good candidate within the window. More precisely, consider the window $g[i, \ldots, i+w]$ for some position i on the genome. The main idea is to find out whether the subsequence $g[i, \ldots, i+h_{\max}]$ contains a good candidate for a half-stem with a corresponding partner half-stem in $g[i+h_{\min}, \ldots, i+w]$. Here, $h_{\min}$ and $h_{\max}$ are lower and upper bounds on the length of the stems we are looking for. To simulate the folding process we align the reversed sequence $g[i+h, \ldots, i]$ against $g[i+h, \ldots, i+w]$, using the Smith-Waterman algorithm for local alignment [31] and a scoring scheme that awards Watson-Crick(A-T, G-C) and wobble base (G-T) pairs and penalizes other pairs and gaps (see Fig. 6 for an illustration of the idea). The default scoring matrix used of all conducted experiments is: +2 for A-T and G-C, +1 for G-U and -1 for all other nucleotide pairs. The algorithm uses linear gap penalties (-1). We also tried more sophisticated scoring based on the actual stacking energies (that are used by RNA folding programs), but it turn out that in this application the simple scoring was superior to the sophisticated one. The default window size and minimal/maximal half stem lengths are 90, 21, and 32, respectively.

Similar approaches have been used in [32] and [33]. They differ in searching for putative stacks ignoring the connecting regions or searching for microRNAs targets instead of microRNA precursor.



(a) (b)

Figure 6: Using local alignment to simulate the folding process (for simplicity, we set $h = h_{\min} = h_{\max}$). (a) Putative microRNA precursor on the genomic sequence g, (b) corresponding alignment.

3.2 Context preservation score

The basic idea behind the *context-preservation-score* (or CPS in short) is the question, how robust a hairpin is relative to its surrounding genomic context: more precisely, if we consider the hairpin, add some genomic context and fold the elongated string again, how much of the actual hairpin is conserved?

Hence, to show that a certain microRNA precursor candidate is robust with respect to its genomic context, we add n nucleotides to both sides of the putative microRNA. For the elongated string we computed the *base pair probability* matrices as proposed by [34]. We did not just consider the MFE structure of the long string, because the base pair probability matrices contains much more information: If the hairpin is not conserved in *the* MFE structure, but in almost all suboptimal structures, we do find the hairpin by considering the base pair probabilities, but we would not find it by just considering the optimal structure.

Then, we take the structure of the actual microRNA (without additional nucleotides) and calculate a probability score as the sum of all probabilities for every interaction in the original candidate structure. We normalize this score by the sequence length to get the CPS value. More formally, the bracket notation of the candidate hairpin of length L consists of a list I of paired nucleotides. Let

$$I = [(i_1, j_1), \ldots, (i_n, j_n)]$$

with (i_k, j_k) representing the interaction between nucleotide i_k and j_k. Then, the context preservation score is given as

$$\mathrm{CPS} = \frac{\sum_{k=0}^{n} P[(i_k, j_k)]}{L}$$

where $P[(i, j)]$ is the probability that the nucleotide i folds onto j as given in the base pair probability matrix that is computed by RNAfold [28].

3.3 Statistical tests

For the random shuffling of nucleotide sequences and calculating p-values for the MFE and the local alignment score of microRNA precursor candidates we use the methods described in [27]. MicroRNA precursor candidate sequences are randomized using dinucleotide shuffling to maintain the same dinucleotide distribution as described by [35].

We use dinucleotide shuffling for randomizing microRNA precursor candidate sequences because it has been shown recently that, relative to mononucleotide shuffling, the preservation of dinucleotide distributions significantly increases the detection of false-positives searching ncRNA elements [36]. This is due the fact that nucleotide stacking is a key determinant of the stability of an RNA secondary structure [35,37]. Additionally, constant dinucleotide composition has been used before successfully to annotate microRNAs in plants [38].

To determine the p-value for the folding minimum free energies as well as for the local alignment scores we use the same randomization tests as in [27] that were used before in several other applications in molecular biology [39]. For all p-value calculations we shuffled each sequence $N = 1000$ times. Statistical tests are then performed in the following manner:

1. Compute the minimum free energy or the local alignment score for the microRNA precursor candidate sequence.

2. Shuffle the original candidate sequence 1000 times using dinucleotide shuffling and compute each time the minimum free energy or the local alignment score to compute the minimum free energy of local alignment score distribution.

3. The p-value is then defined as

$$p = \frac{R}{N+1} \ ,$$

where N is the number of iterations and R the number of shuffled sequences that have a minimum free energy less or equal the original minimum free energy or a local alignment score above or equal the original alignment score.

3.4 Conservation graph

MicroRNA sequences are often strongly evolutionary conserved. To compare k different genomes we first have to apply miRo to each genome separately to get k candidate lists. Every half stem of each microRNA precursor candidate is then aligned against every half stem of each other candidate from the other genomes, taking care of all different possibilities of orientation, so that we are independent from the transcription sense. We compute a quadratic number of local alignments to detect these similarities using again the Smith-Waterman algorithm, but now with a standard alignment scoring scheme. Then, we define an undirected *conservation graph* $G_c = (V_c, E_c)$ as follows (Figure 5):

- Each node in V_c corresponds to a microRNA precursor candidate.
- An edge $(c_1, c_2) \in E_c$ connects two candidates if and only if the score of a local alignment between different half-stems of two candidates is larger than a given threshold.

We now enumerate all maximal cliques in G_c of minimal size m using the Bron-Kerbosch algorithm [40]. A clique in G_c is a set of vertices V_c such that for every two vertices in V_c, there exists an edge connecting the two. This is equivalent to saying that the subgraph induced by V_c is a complete graph. The size of a clique is given by the number of vertices it contains. The user-defined parameter m determines the minimal number of genomes a microRNA precursor candidate has to be conserved in. Each clique C corresponds to a conserved microRNA precursor candidate in $|C|$ genomes, where $|C|$ denotes the size of clique C.

4 Conclusions

We have presented miRo, a hierarchical pipeline for microRNA precursor detection. Whereas most other methods need additional information in form of multiple alignments or previously known microRNA sequences, our approach detects putative microRNA precursors by analysing the precursor structure alone. Additional knowledge on the candidates or their targets can however easily be integrated into our open and flexible software framework. We have implemented miRo in C^{++} as part of the freely available open software library LiSA[1].

Several case studies with miRo demonstrate that our new alignment method for microRNA precursor has a comparable discriminative power as methods based on folding. miRo can successfully identify promising microRNA candidates in reasonable computation time. We could show that our alignment scores are similarly significant for microRNA precursor structure than MFE values, and we introduced a new context-preservation score which is able to distinguish between secondary structures that are conserved within their genomic context and non- conserved candidates. For the Epstein-Barr virus genome we detected all known microRNAs from miRBase database, and we found several new promising microRNA candidates. Additionally, we

[1] www.planet-lisa.net

propose a comparative method based on graph-theory to include, if reasonable, evolutionary information into our prediction scheme. Testing our comparative method we found several conserved candidates in between different herpesvirus genomes.

References

[1] Mattick JS: **Challenging the dogma: the hidden layer of non-protein RNAs in complex organisms**. *Bioessays* 2003, **25**:930–939.

[2] Mattick JS: **RNA regulation: a new genetics?** *Nature Genetics* 2004, **5**:316–323.

[3] Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR: **Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.** *Genome Res* 2004, **14**(3):331–342.

[4] Johnson JM, Edwards S, Shoemaker D, Schadt EE: **Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments**. *Trends Genet.* 2005, **21**(2):93–102.

[5] Imanishi T, *et al*: **Integrative annotation of 21,037 human genes validated by full-length cDNA clones.** *PLoS Biol* 2004, **2**(6):e162.

[6] Cummins JM, He Y, Leary RJ, Pagliarini R, Diaz LA, Sjoblom T, Barad O, Bentwich Z, Szafranska AE, Labourier E, Raymond CK, Roberts BS, Juhl H, Kinzler KW, Vogelstein B, Velculescu VE: **The colorectal microRNAome.** *Proc Natl Acad Sci U S A* 2006, **103**(10):3687–3692.

[7] Washietl S, Hofacker IL, Lukasser M, Huttenhofer A, Stadler PF: **Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome**. *Nat Biotechnol.* 2005, **23**(11):1383–1390.

[8] Missal K, Rose D, Stadler PF: **Non-coding RNAs in *Ciona intestinalis***. *Bioinformatics* 2005, **21**(S2):i77–i78.

[9] Missal K, Zhu X, Rose D, Deng W, SkogerbøG, Chen R, Stadler PF: **Prediction of structured non-coding RNAs in the genomes of the nematodes Caenorhabditis elegans and Caenorhabditis briggsae.** *J Exp Zoolog B Mol Dev Evol* 2006, **306**(4):379–392.

[10] Ambros V: **The functions of animal microRNAs**. *Nature* 2004, **431**:350–355.

[11] Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function**. *Cell* 2004, **116**:281–297.

[12] Sullivan CS, Ganem D: **MicroRNAs and viral infection**. *Cell* 2005, **20**:3–7.

[13] He L, Hannon G: **MicroRNAs: small RNAs with a big role in gene regulation**. *Nat. Rev. Genet.* 2004, **5**:522–531.

[14] Washietl S, Hofacker IL, Stadler PF: **Fast and reliable prediction of noncoding RNAs**. *PNAS* 2005, **102**(7):2454–2459.

[15] Rivas E, Eddy SR: **Noncoding RNA gene detection using comparative sequence analysis.** *BMC Bioinformatics* 2001, **2**:8.

[16] Zhang S, Haas B, Eskin E, Bafna V: **Searching Genomes for Noncoding RNA Using FastR**. *IEEE/ACM Trans. Comput. Biology Bioinform.* 2005, **2**(4):366–379.

[17] Klein RJ, Eddy SR: **RSEARCH: finding homologs of single structured RNA sequences.** *BMC Bioinformatics* 2003, **4**:44.

[18] Hertel J, Stadler PF: **Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data**. *Bioinformatics* 2006, **22**(14):e197–202.

[19] Dezulian T, Remmert M, Palatnik J, Weigel D, Huson D: **Identification of plant microRNA homologs**. *Bioinformatics* 2006, **22**(3):359–360.

[20] Wang X, Zhang J, Li F, Gu J, He T, Zhang X, Li Y: **MicroRNA identification based on sequence and structure alignment**. *Bioinformatics* 2005, **21**(18):3610–3614.

[21] Sewer A, Paul N, Landgraf P, Aravin A, Pfeffer S, Brownstein M, Tuschl T, van Nimwegen E, Zavolan M: **Identification of clustered microRNAs using an ab initio prediction method**. *BMC Bioinformatics* 2005, **6**:267.

[22] Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: **miRBase: microRNA sequences, targets and gene nomenclature**. *Nucleic Acids Research* 2006, **34**(Database Issue):D140–D144.

[23] Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffith-Jones S, Marshall M: **A uniform system for microRNA annotation**. *RNA* 2003, **9**:277–279.

[24] Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, Kim V: **The nuclear RNase III Drosha initiates mircoRNA processing**. *Nature* 2003, **25**(425):415–419.

[25] Hutvagner G, McLachlan J, Pasquinelli A, Balint A, Tuschl T, Zamore P: **A cellur function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA**. *Science* 2001, **293**(5531):834–838.

[26] Lund E, Guttinger S, Dahlberg AC, Kutay U: **Nuclear export of microRNA precursors**. *Science* 2004, **303**(5654):95–98.

[27] Bonnet E, Wuyts J, Rouze P, Van de Peer Y: **Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences**. *Bioinformatics* 2004, **20**(17):2911–2917.

[28] Hofacker IL: **Vienna RNA secondary structure server**. *Nucl. Acids Res.* 2003, **31**(13):3429–3431.

[29] Tanzer A, Amemiya CT, Kim CB, Stadler PF: **Evolution of MicroRNAs Located Within Hox Gene Clusters**. *J.Exp.Zool.* 2005, **304B**:75–85.

[30] Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E, Sharon E, Spector Y, Bentwich Z: **Identification of hundreds of conserved and nonconserved human microRNAs**. *Nature Genetics* 2005, **37**(7):766–770.

[31] Smith TF, Waterman MS: **Identification of Common Molecular Subsequences**. *J. Mol. Biology* 1981, **147**:195–197.

[32] Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS: **MicroRNA targets in *Drosophila***. *Genome Biology* 2003, **5**:R1.1–R1.14.

[33] Bafna V, Tang H, Zhang S: **Consensus Folding of Unaligned RNA Sequences Revisited**. *J. Comput. Biol.* 2006, **13**(2):283–295.

[34] McCaskill JS: **The Equilibrium Partition Function and Base Pair Binding Probabilities for RNA Secondary Structure**. *Biopolymers* 1990, **29**:1105–1119.

[35] Workman C, Krogh A: **No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution**. *Nucleic Acids Res.* 1999, **27**:4816–4822.

[36] Babak T, Blencowe B, Hughes T: **Considerations in the identification of functional RNA structural elements in genomic alignments.** *BMC Bioinformatics* 2007, **8**:33.

[37] Uzilov AV, Keegan JM, Mathews DH: **Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change.** *BMC Bioinformatics* 2006, **7**:173.

[38] Bonnet E, Wuyts J, Rouze P, Van de Peer Y: **Detection of 71 potential conserved plant microRNAs in *Arabidopsis thaliana* and *Oryza sativa* identifies important target genes**. *PNAS* 2004, **20**(17):2911–2917.

[39] Manly BF: *Randomization, Bootstrap and Monte Carlo Methods in Biology, Volume Chapman Hall.* Chapman & Hall/CRC; 1997.

[40] Bron C, Kerbosch J: **Algorithm 457 - finding all maximum cliques of an undirected graph**. *Comm. ACM* 1973, **16**:575–577.