BIOINFORMATICS

Vol. 26 ECCB 2010, pages i426–i432
doi:10.1093/bioinformatics/btq363

Classification of ncRNAs using position and size information in deep sequencing data

Florian Erhard* and Ralf Zimmer

Institut für Informatik, Ludwig-Maximilians-Universität München, Amalienstraße 17, 80333 München, Germany

ABSTRACT

Motivation: Small non-coding RNAs (ncRNAs) play important roles in various cellular functions in all clades of life. With next-generation sequencing techniques, it has become possible to study ncRNAs in a high-throughput manner and by using specialized algorithms ncRNA classes such as miRNAs can be detected in deep sequencing data. Typically, such methods are targeted to a certain class of ncRNA. Many methods rely on RNA secondary structure prediction, which is not always accurate and not all ncRNA classes are characterized by a common secondary structure. Unbiased classification methods for ncRNAs could be important to improve accuracy and to detect new ncRNA classes in sequencing data.

Results: Here, we present a scoring system called ALPS (alignment of pattern matrices score) that only uses primary information from a deep sequencing experiment, i.e. the relative positions and lengths of reads, to classify ncRNAs. ALPS makes no further assumptions, e.g. about common structural properties in the ncRNA class and is nevertheless able to identify ncRNA classes with high accuracy. Since ALPS is not designed to recognize a certain class of ncRNA, it can be used to detect novel ncRNA classes, as long as these unknown ncRNAs have a characteristic pattern of deep sequencing read lengths and positions. We evaluate our scoring system on publicly available deep sequencing data and show that it is able to classify known ncRNAs with high sensitivity and specificity.

Availability: Calculated pattern matrices of the datasets hESC and EB are available at the project web site http://www.bio.ifi.lmu.de/ALPS. An implementation of the described method is available upon request from the authors.

Contact: florian.erhard@bio.ifi.lmu.de

1 INTRODUCTION

Next-generation sequencing platforms such as Solexa/Illumina, ABI SOLiD or 454/Roche are extensively used to sequence small RNAs of roughly 14–36 nt length at astonishing rates in various organisms (Babiarz *et al.*, 2008; Czech *et al.*, 2008; Kato *et al.*, 2009; Morin *et al.*, 2008; Rathjen *et al.*, 2009). For instance, they are used to determine expression profiles of miRNAs, 20–24 nt long RNA molecules, that have emerged in recent years as important post-transcriptional regulators in all known multicellular organisms and that are known to play roles in development, tumorigenesis and viral infection (Bartel, 2004). Besides miRNAs other small non-coding RNA (ncRNA) classes such as piRNAs (Aravin *et al.*, 2001), snoRNAs (Bachellerie *et al.*, 2002) or scaRNAs (Gerard *et al.*, 2010) have been investigated. Only recently, 454 sequencing revealed the existence of 16 nt long RNA (therefore termed unusual small RNA or usRNAs) in cells infected with KSHV (Li *et al.*, 2009). usRNAs are derived from both virus and host cell and are associated with the RNA-induced silencing complex (RISC). Advances in throughput, accuracy and the ability to sequence longer reads will not only lead to more and more precise detection of already known ncRNA classes, but also to the discovery of new types. It is therefore of great interest to develop methods for automatic classification of ncRNA using deep sequencing data.

Most ncRNAs have very specific structural properties that have been used to classify them (Will *et al.*, 2007), e.g. tRNAs possess a cloverleaf structure, whereas miRNA precursors form stable hairpins. However, these methods rely on the prediction of RNA secondary structure and even for short molecules the current RNA secondary structure energy model is not always able to predict the native structure (Doshi *et al.*, 2004; Dowell and Eddy, 2004). This can be due to the incompleteness of the nearest neighbor energy model or explained by the fact that the minimal free energy structure is not necessarily the native one due to unknown modifications or effects of folding kinetics (Higgs and Morgan, 1995). In the case of *de novo* prediction of, e.g. miRNAs, the exact pre-miRNA sequence is not known a priori. Even if the hairpin can be predicted for the pre-miRNA sequence, it could be disrupted, if a few bases upstream or downstream are appended or removed from this sequence. Therefore, multiple windows around a putative miRNA are folded or a local folding tool such as RNALfold (Hofacker *et al.*, 2004) is used. This however necessarily leads to an increased false positive rate since many genomic sequences that do not encode miRNAs can fold into stable hairpins (Bentwich, 2005).

Deep sequencing offers additional criteria to distinguish ncRNA classes. A typical experimental setup is to determine the content of small ncRNA in a cell under certain conditions. Therefore, only intervals on the genome are considered, where enough sequencing reads have been aligned to. The specific number of reads depends on the tradeoff between sensitivity and specificity. If the experiment aims to identify a special class of ncRNAs, specialized algorithms can be applied that detect specific features of that ncRNA class based on biological knowledge. For example, in miRNA biogenesis, one strand of the precursor is preferentially included into RISC (the mature miRNA) and the other is rapidly degraded (miRNA star). Considering this bias together with structural miRNA properties can dramatically increase specificity of miRNA detection, as shown before (Friedlander *et al.*, 2008; Morin *et al.*, 2008).

miRNAs recognize their targets by their seed region (Grimson *et al.*, 2007) and, due to their biogenesis, have specific lengths (MacRae *et al.*, 2007). Both features of miRNAs should be detectable in an excess of deep sequencing reads that align to a specific genomic position and have a specific length. However, this is not always the case for miRNAs in large-scale experiments (e.g. Morin *et al.*, 2008). The read start position of many miRNAs follow a narrow distribution that is often skewed toward the miRNA 3′ end

*To whom correspondence should be addressed.

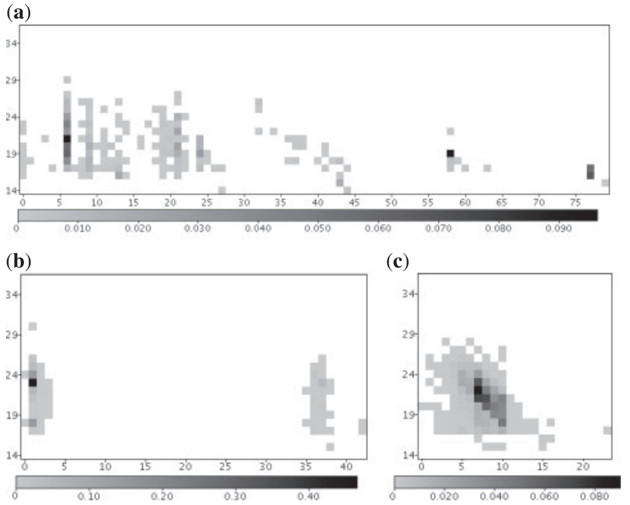

Fig. 1. Typical position and length-dependent pattern matrices for (**a**) a tRNA (Gly/GCC), (**b**) a miRNA (mir-423) and (**c**) a snRNA (U2). Frequencies of reads starting at position (*x*-axis) and of length (*y*-axis) are visualized in different shades of gray. Note that both the snRNA and the tRNA could easily be mistaken for a miRNA, if only the most abundant read is considered. Graphical respresentations for all pattern matrices are available on the project web site.

and read lengths are often variable (see also Fig. 1b). Such alternative mature miRNA forms are often referred to as isomiRs (Morin *et al.*, 2008).

In addition to positioning and lengths of reads, distances of reads aligned in close proximity of other reads also carry information about ncRNA classes: at least for animals, the miRNA star should be detectable at a distance of roughly 40 nt to the mature miRNA (Friedlander *et al.*, 2008). Distance information also helps to distinguish miRNAs from degradation products of other abundant RNA species such as tRNAs or snRNAs (Fig. 1). And, most importantly, using this information can help to classify novel ncRNA or ncRNAs that do not possess a characteristic secondary structure.

In this article, we show how to exploit position and length-dependent read patterns to classify ncRNAs. We make no further assumptions about structural and other class-specific properties and only consider primary information from the alignment of deep sequencing reads on the genome. Our method ALPS (alignment of pattern matrices score) allows to detect miRNAs and other known ncRNA classes with high accuracy and due to its unbiased nature, it also provides a straight-forward way to discover and classify novel ncRNAs. Our approach is complementary to existing methods that rely on structural properties and we expect that their combination with our approach allows to increase their sensitivity and specificity.

2 APPROACH

The starting point for ALPS is the output of a short-read aligner [e.g. Bowtie (Langmead *et al.*, 2009) or BWA (Li *et al.*, 2009)] consisting of the positions in the genome where deep sequencing reads have been aligned to. Then, intervals are identified by clustering these positions such that (i) each interval contains at least m reads, (ii) there is no consecutive part of length $>t$ within an interval, that is not covered by a read; and (iii) t nucleotides downstream and upstream are not covered by a read. The classification problem of ncRNAs using deep sequencing data then is to assign a class label, e.g. *miRNA,tRNA,snoRNA, etc*., to each of these intervals. For a well-annotated organism such as human, mouse or yeast such class labels are already available for many of these intervals in public databases. Then, class labels for the intervals without annotation can be predicted based on similarity to intervals with known annotation, which is often called (semi-) supervised learning. If no or only very few annotations are available for the organism in question, intervals can still be clustered in an unsupervised manner. Both approaches need a way to calculate the similarity between two intervals.

ALPS is such a similarity score computed by an alignment of their so-called pattern matrices. These contain the information about the positions and lengths of aligned reads. Since we cannot assume that all exact distances between aligned reads are always representative for an ncRNA class, we allow gaps in ALPS. For instance, to respect the distance of the mature miRNA and their corresponding miRNA star, our algorithm must be allowed to align the start positions of the two mature miRNAs as well as the start positions of the two miRNA stars, even if the loops of the two precursor miRNAs have different lengths.

Usually, for many intervals, annotations are already available in public databases and these can be used to classify unknown—so far not annotated—intervals similar to them. Generally, ALPS similarities are not biased toward a special class of ncRNAs since they are only based on the primary data from the deep sequencing experiment. Therefore, used as a distance measure for any unsupervised clustering, the similarity of pattern matrices score will find groups of ncRNAs, that exhibit similar distributions (with respect to relative position and length) of deep sequencing reads. If such a distribution is characteristic for an unknown class of ncRNAs, the clustering based on our score should be able to detect it.

In this article, the focus is not on the detection of unknown classes and hierarchies of ncRNAs but on the detection of already known ncRNA classes to demonstrate the usefulness of our scoring system. Based on annotations retrieved from mirBase (Griffiths-Jones *et al.*, 2008), gtRNAdb (Chan and Lowe, 2009), Ensembl and Refseq, we identify intervals of known ncRNA classes in published deep sequencing data and benchmark our scoring system based on its ability to reassign an interval to its correct class, after its class label has been removed.

3 METHODS

To identify the set of intervals $\mathcal{I}$ and their corresponding pattern matrices, we iterate over the sorted read alignments and add a read $r=(r_1,r_2)$ to the current interval $I=(i_1,i_2)$ as long as $i_2>r_1-t$, where r_1 and r_2 are genomic start and end of r, respectively, and t is a user-defined tolerance (we use $t=50$ throughout the article). Since we do these iterations per chromosome and per strand, each interval spans reads that mapped to one strand of a single chromosome in close proximity to each other and reads of two different intervals are either on different strands or chromosomes or more than t nt apart from each other. An entry $N^I[l,i]$ of the *pattern matrix* N^I of interval I is the number of reads of length l starting at position i in this interval. Positions are according to the strand direction, i.e. if i_1 and i_2 are genomic start and end of an interval on the−strand and a read $r=(r_1,r_2)$ falls into that interval, it contributes to the entry $N^I[r_2-r_1,i_2-r_2]$ of the pattern matrix. Since we want to compare pattern matrices for similarity regarding bias of read start positions and lengths frequencies and we have to respect that two ncRNAs of the same class can be expressed at different levels, we normalize each pattern matrix:

$$\tilde{N}^I[l,i]=\frac{N^I[l,i]}{\sum_{l',i'}N^I[l',i']} \qquad (1)$$

To quantify the similarity of two intervals $I,J\in\mathcal{I}$, we consider their normalized pattern matrices $\tilde{N}^I$ and $\tilde{N}^J$ as sequences of column vectors $(\tilde{N}^I[\bullet,i])_{i=1..|I|}$ and $(\tilde{N}^J[\bullet,j])_{j=1..|J|}$ and compute their optimal alignment. Here we adopt the notation, that, $A[\bullet,i]$ is the i-th column vector of matrix A. Thus, a column vector is the length distribution of deep sequencing reads that start at a certain position within the interval. Note that this distribution is normalized to the proportion of reads that start at this position. The similarity score $S^{I,J}(i,j)$ for aligning position i in interval I to position j in interval J is computed according to

$$S^{I,J}(i,j)=(\tilde{N}^I[\bullet,i])^{\mathrm{T}}\otimes M\otimes\tilde{N}^J[\bullet,j] \qquad (2)$$

where M is a $L\times L$ matrix (L is the maximal read length). In the simplest case, the identity matrix $M=id_L$ is used and $\otimes$ is the usual matrix multiplication. Then the similarity score is basically just the scalar product of the corresponding column vectors. However, since ncRNA classes are usually not defined by a specific length but by a narrow distribution of lengths, it is reasonable to reward not only exact length matches but also small differences and to penalize large deviations of peaks in the length distributions. Therefore, we use a matrix $M=H^{k,\lambda}$ derived from the sigmoidal function:

$$H[i,j]^{k,\lambda} = h^{k,\lambda}(|i-j|) \qquad (3)$$

$$h^{k,\lambda}(x) = 1-\frac{2x^k}{\lambda^k+x^k} \qquad (4)$$

This matrix rewards differences in read lengths, as long as the absolute difference is at most λ and penalizes all deviations of more than λ. The parameter k describes the steepness of rewards and penalties. The standard sum-product matrix multiplication can also be replaced by a sum-min matrix multiplication. If $M=id_L$ is used and the two column vectors are considered as functions, this score can be geometrically interpreted as their common integral. Again, a hill function derived matrix $H^{k,\lambda}$ can be used to respect length distributions (after negative entries in the matrix have been removed). The ALPS similarity, i.e. the optimal alignment score of the two intervals I and J then is:

$$\hat{s}(I,J) = \max_A\left\{\sum_{(i,j)\in A}S^{I,J}(i,j)+\sum_{n\in G(A)}g(n)\right\} \qquad (5)$$

$$g(n) = o+e\cdot n \qquad (6)$$

The maximum in Equation (5) is over all possible alignments A of the intervals I and J and $G(A)$ is the set of all gaps in alignment A. Note that the affine gap cost function (6) penalizes many short gaps more than few long gaps, which is important for our similarity scoring. We can calculate $\hat{s}(I,J)$ efficiently using the algorithm of Gotoh (1982) in time $\mathcal{O}(|I|\cdot|J|\cdot L)$ after a preprocessing of the scoring function S in time $\mathcal{O}(|J|\cdot L^2)$. The preprocessing involves the computations of the second matrix multiplication $M\otimes\tilde{N}^J[\bullet,j]$ for all $j\in[1;|J|]$.

The score in Equation (5) corresponds to an optimal *global* alignment. However, we can also define other variants of ALPS similarity: the optimal *freeshift* (also often called *semi-global*) alignment score $\hat{s}^f(I,J)$ is given as in Equation (5) by replacing $G(A)$ by $G^f(A)$ that contains all gaps from $G(A)$ but the longer of the two leading gaps and the longer of the two trailing gaps. Similarly, for the optimal *local* alignment score, $\hat{s}^l(I,J)$, $G^l(A)$ is used instead of $G(A)$, that contains all but both leading and both trailing gaps. This is equivalent to the usual definition of local alignment, i.e. the optimal global alignment of two subsequences. Note that we can compute the optimal local and freeshift alignments efficiently using a modified version of the Gotoh algorithm, as suggested by Smith and Waterman (1981).

Thus, a scoring system for pairwise ALPS similarities can be described by the 5-tuple $\mathcal{S}=(M,\otimes,o,e,mode)$, where M is the matrix and $\otimes$ the operator for the calculation of the column vector similarity, respectively, o,e are the gap open and gap extend parameters for the affine gap cost function and *mode* is the alignment mode (global, local or freeshift).

Table 1. Annotations from mirBase, gtRNAdb, Ensembl and RefSeq, ordered by their priority used for the initial class assignment

Origin	Annotation	Combined	hESC	EB
mirBase/Ensembl	miRNA	miRNA	103	101
Ensembl	miRNA_pseudogene	miRNA		
gtRNAdb/Ensembl	tRNA	tRNA	158	99
Ensembl	tRNA_pseudogene	tRNA		
Ensembl	Mt_tRNA	tRNA		
Ensembl	Mt_tRNA_pseudogene	tRNA		
Ensembl	rRNA	rRNA	43	27
Ensembl	rRNA_pseudogene	rRNA		
Ensembl	Mt_rRNA	rRNA		
Ensembl	snRNA	snRNA	13	12
Ensembl	snRNA_pseudogene	snRNA		
Ensembl	snoRNA	snoRNA	10	6
Ensembl	snoRNA_pseudogene	snoRNA		
Ensembl	misc_RNA	misc_RNA	94	85
Ensembl	misc_RNA_pseudogene	misc_RNA		
Ensembl	lincRNA	misc_RNA		
Ensembl	scRNA	misc_RNA		
Ensembl	scRNA_pseudogene	misc_RNA		
Ensembl	Pseudogene	misc_RNA		
RefSeq	CDS	misc_RNA		
RefSeq	INTRON	misc_RNA		
RefSeq	UTR	misc_RNA		
RefSeq	3FLANK	misc_RNA		
RefSeq	5FLANK	misc_RNA		
	Unknown	Unknown	80	56

Similar annotations are combined and the number of respective intervals in the two datasets used for benchmarking is given.

We compute the pairwise ALPS similarities $\hat{s}(I,J)$ for all intervals $I,J\in\mathcal{I}^{m,t}$ that contain at least m reads with tolerance t given a scoring system $\mathcal{S}$. Then we assign a class to each of the intervals by using annotations from mirBase (Griffiths-Jones *et al.*, 2008), gtRNAdb (Chan and Lowe, 2009), Ensembl and RefSeq. For intervals with multiple assigned annotations, we prioritize annotations according to Table 1 and we combine similar annotations. All intervals annotated with A are thus partitioned into a cluster C^A. We define the inner and outer similarity scores of class A as the sets

$$D^{\text{inner}}(A) = \{\hat{s}(I,J)|I,J\in C^A\} \qquad (7)$$

$$D^{\text{outer}}(A) = \{\hat{s}(I,J)|I\in C^A,J\notin C^A\} \qquad (8)$$

Using their respective distributions $P^{\text{inner}}(A)$ and $P^{\text{outer}}(A)$, we can estimate the ability of $\mathcal{S}$ to separate A from all other classes. This means, by using general optimization techniques such as simple grid search, genetic algorithms or specialized methods such as VALP (Zien *et al.*, 2000), we can optimize $\mathcal{S}$ for many purposes, e.g. a median-based hierarchical clustering that is supposed to separate all classes equally well would require a scoring system, that maximizes $\sum_A\text{median}(P^{\text{inner}})-\text{median}(P^{\text{outer}})$.

Here, we use only $P^{\text{outer}}(A)$ to test the null hypothesis that an interval I without annotation is not from class A. We calculate an empirical P-value for each $\hat{s}(I,J),J\in C^A$ from the right tail of $P^{\text{outer}}(A)$ and combine each of these $|C^A|$ P-values using Fisher's method (Fisher, 1970). We then select the class with the smallest P-value.

4 RESULTS

We applied our method to previously published Illumina sequencing data (Morin *et al.*, 2008), where small RNAs of human embryonic

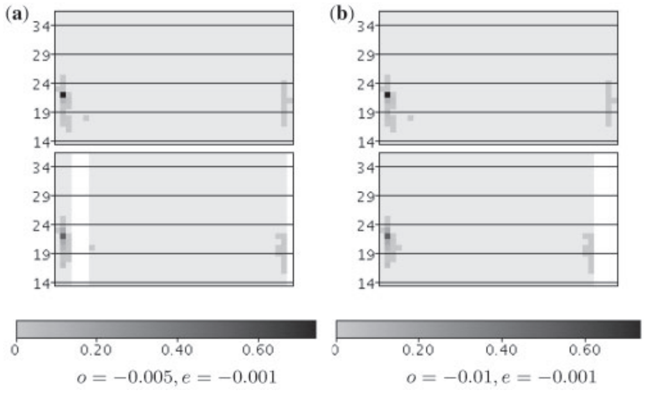

Fig. 2. Freeshift alignments of hsa-mir-99b (top) and hsa-mir-185 (bottom). Pattern matrix frequencies are visualized in different shades of gray and white areas correspond to unaligned parts of the matrices. Note that with the gap cost function in 2(a), the miRNA star start positions on the right parts of the matrices are correctly aligned, whereas slightly altered parameters in 2(b) erroneously move the necessary gap to the end, where it is not penalized.

stem cells (hESC) and embryoid body cells (EB) have been sequenced. We used Bowtie to align the trimmed reads to the human genome (hg19) obtained from the UCSC genome browser. We allowed no mismatches but did not restrict the number of loci a read can be aligned to. We identified intervals as described in Section 3 ($t=50, m=1000$) and assigned them to the classes in Table 1. We determined the normalized pattern matrices (see project web site for graphical visualizations) and computed all pairwise ALPS similarities for various scoring systems.

First, we checked which choices of gap parameters make differences in the alignments of intervals. We considered the intervals of hsa-mir-99b and hsa-mir-185 that are both 5′ donors (i.e. the mature miRNA originates from the 5′ arm of the precursor), are expressed at similar levels (1892 and 2148 reads in EB, respectively) and have different loop lengths. Thus, a correct alignment must introduce a gap between the positions of the mature miRNA and the miRNA star in the sequence of column vectors of mir-185 (which has the shorter loop). If we calculate the optimal freeshift alignment using the Hill matrix $H^{2,1}$ and min-product matrix multiplication, gap parameters of $o=-0.005$ and $e=-0.001$ are indeed able to produce a correct alignment (Fig. 2). We emphasize, that meaningful ranges of gap parameters are highly dependent on the other parameters and that automated parameter optimization could resolve these ranges.

A second theoretical consideration can be made by examining the inner and outer score distributions (Fig. 3). When the sum-min and the sum-product operator is used with the same matrix (the identity matrix), scores of the former naturally tend to be higher than scores of the later. If the identity matrix is replaced by $H^{4,1}$ or $H^{8,1}$, scores also tend to increase. For all parameter choices, it is apparent that inner and outer scores are significantly different, but their distributions are not completely separated. The outer distribution describes all ALPS similarities between pairs of intervals, one annotated as miRNA, the other not annotated as miRNA. However, mirBase is not complete, and as a consequence, it is possible that the outer score distribution contains miRNA–miRNA scores, which can explain the elongated right tail of all P^{outer}. The inner distribution consists of all pairwise ALPS similarities of two intervals both annotated as miRNA. Especially when using $M=id_L$, many scores tend to be small, since only exact agreements in length are rewarded, and two mature miRNAs may have differing sizes. In addition, miRNA may be 5′ donors or 3′ donors or both mature and miRNA stars are expressed at very similar levels. As a consequence, P^{inner} does not only contain overall high scores, but also scores indicating differing subclasses.

However, all these parameter choices are able to separate miRNAs from other ncRNAs, when we use all scores for classification. Using any aggregate statistics fails in many cases: if the maximal scores is used, a true miRNA may be too similar to an interval with unknown annotation, which is in fact a still unknown miRNA, leading to a misclassification. If one uses the minimum, the inner scoring is hampered by subclasses. Therefore, using all scores and a statistically robust method to combine them (such as Fisher's method) is necessary for reliable classification.

In order to assess whether ALPS is able to classify ncRNA reliably, we applied the following procedure: each annotated interval I was removed from its cluster C^A and the described method was used to determine the class of I. Since we did not restrict the number of loci a read could align to, and many of the abundant ncRNAs are present in multiple copies in the human genome, we considered only scores $\hat{s}(I,J)$ where the genomic sequences of I and J did not contain common subsequences of length >10, i.e. no deep sequencing read has been counted in both intervals I and J. For all other scores, P-values were calculated and combined as described. We then calculated recall and precision for each class A separately as the number of intervals correctly assigned divided by the number of intervals originally belonging to C^A (recall) and divided by the number of intervals assigned to C^A (precision), respectively.

As indicated above, we tried various parameter combinations to classify ncRNAs. Since there are only very few unique snRNAs, snoRNAs and rRNAs, we only considered miRNAs and tRNAs for evaluation. Except for some obviously too extreme parameter combinations (e.g. too negative gap parameters for global alignments), the classification performance was remarkably stable with recall values of up to 98% at a precision of 60% for miRNAs (Fig. 4). These relatively low precision values in the miRNA class rise the question, whether our scoring tends to classify too many intervals as miRNAs. However, the classes *unknown* and *misc_RNA* are not excluded from our analyses, and nearly all of the intervals additionally assigned to the class miRNA originate from *unknown* and *misc_RNA* whose pattern matrix indeed is very similar to that of miRNAs. We predicted the secondary structures of the corresponding sequences using RNAfold (Hofacker *et al.*, 1994) and some of them are indeed predicted to be able to fold into hairpins. Whether these reads really correspond to mature miRNAs, are degradation products or otherwise processed RNAs must still be elucidated, however.

Here, we applied our method only to abundant ncRNAs. This is inherent to the method as we have to estimate the distribution of read lengths per position for an ncRNA gene, which is only possible, if enough reads have been sequenced. Due to further development of current sequencing techniques, it will be possible to achieve more and more sequencing depth at lower costs and therefore, also low abundant ncRNAs will be represented by enough sequencing reads.

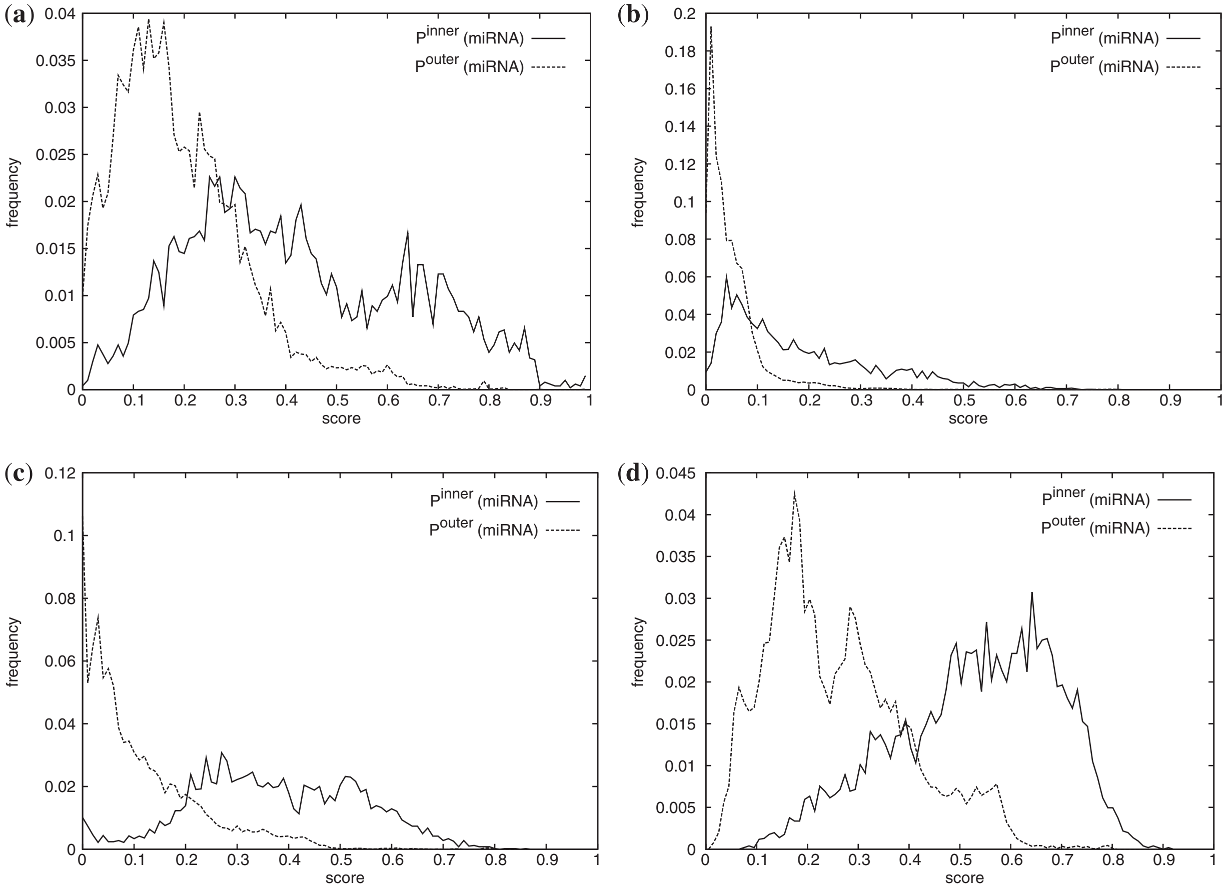

Fig. 3. Inner and outer score distributions for miRNAs. The EB dataset is used and shown are scores for freeshift alignments with $o = -0.05$, $e = -0.01$. Matrices and operators for the calculation of column vector similarities are (**a**) $M = id_L, \otimes =$sum-min, (**b**) $M = id_L, \otimes =$sum-product, (**c**) $M = H^{4,1}, \otimes =$sum-product and (**d**) $M = H^{8,1}, \otimes =$sum-product, respectively. In all cases, the inner and outer distributions are significantly different, although not completely separated.

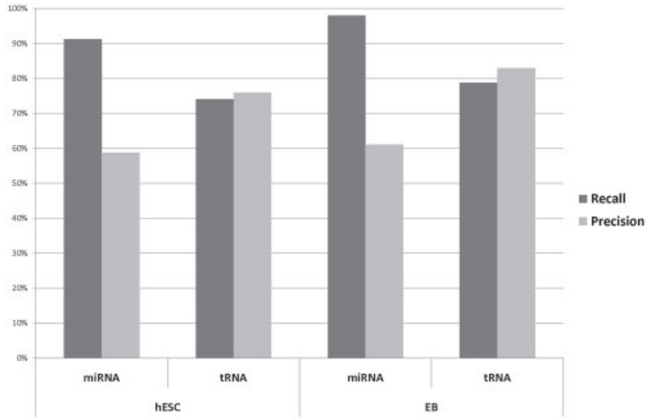

Fig. 4. Precision and recall values for the scoring system ($H^{2,1}$, sum-min, -0.01, -0.005, freeshift).

5 DISCUSSION

Deep sequencing reads of ncRNAs follow very specific patterns regarding their length and positions with respect to their genes. Classes of ncRNAs are defined by their function and biogenesis and often share a common structure. Each of these can contribute to a biased distribution of reads on the ncRNA gene:

- Regulatory RNAs such as miRNAs, piRNAs or siRNAs are believed to recognize their targets by a short complementary region (seed). Therefore, for a proper function, the cell has to take care that the seed of these RNA classes is not shifted and, as a consequence, a wealth of deep sequencing reads starting at specific positions should be detectable. This can be observed in the pattern matrices computed for high-throughput sequencing data. The consideration of reads starting at adjacent positions allows to distinguish these ncRNA classes from degradation products of other abundant species.
- The specific pattern observed for longer ncRNAs such as tRNAs (see Fig. 1a) can possibly be explained by their degradation: cleavage by RNAses can be biased toward certain parts of the tRNA, which leads in the case of the cloverleaf structured tRNAs to a pattern of tRNA halves or quarters (Thompson and Parker, 2009). Although some of these degradation products can be mistaken for, e.g. a miRNA due to similar length, the consideration of longer intervals and the

distances between such subintervals can be used to separate these classes of ncRNAs. It has been observed, that degradation products of tRNAs are associated with RISC (Haussecker *et al.*, 2010), which could explain miRNA-like read patterns of tRNAs. In spite of that ALPS is able to separate these tRNAs from miRNAs.

- Patterns generated by miRNA biogenesis are obvious when looking at graphical representations of pattern matrices. In addition to the mature miRNA and the miRNA star, additional reads are present for some intervals. These can either be explained by degradation products or by additional drosha products that have been observed previously (Shi *et al.*, 2009).

It is currently unclear which classes of ncRNAs exhibit characteristic patterns of deep sequencing reads, but the points discussed above indicate that in theory all ncRNA classes defined by a common function or biogenesis should have such a pattern and should therefore be amenable to classification by ALPS.

As indicated in Figure 2, exact distances between the start positions, e.g. of the mature miRNA and the miRNA star within such patterns are not fixed and in plants, the miRNA hairpin is longer and even more variable than in animals. Allowing gaps in the alignment, therefore, enables ALPS to compute reasonable similarity scores for ncRNA classes, where such distances are highly variable. It is furthermore important to allow affine gap cost functions since linear gap costs tend to disrupt correct alignments. Gap parameters can be adjusted, such that single alignments become correct (e.g. as in Fig. 2). However, we observed that classification accuracy in our test datasets is not heavily influenced by gap parameters. This is a consequence of the strong signal of the mature miRNA read that contributes in many cases enough score to the ALPS similarity to separate miRNAs from non-miRNAs. Thus, for classification of ncRNAs exhibiting such dissimilar patterns as in our test set, results are very robust and independent of the scoring system, i.e. a single reasonable scoring system can be used for classification of all ncRNAs in such a case.

If patterns for ncRNA classes are not as distinct as for the miRNAs and tRNAs in our test set, gap parameters and the matrix M can be tuned for proper classification. We described an approach to evaluate parameter sets based on the inner and outer score distributions and we note that already available methods to optimize other alignment-based scoring systems, e.g. for homology modeling and of protein (structure) alignment can directly be applied to ALPS.

We emphasize, that ALPS similarities should be calculated per experiment and comparisons across different datasets, generated in different labs with different protocols or even different sequencing platforms, should be performed with care. In the two datasets, that we used for validation, pattern matrices were highly concordant for intervals observed in both datasets. However, it is not clear how much technical bias is introduced into pattern matrices, i.e. pattern components that are not due to biology but introduced by technical factors. Comparing pattern matrices of different protocols or sequencing techniques is subject for further studies.

6 CONCLUSION AND OUTLOOK

We developed an alignment-based method that allows to quantify the similarity of ncRNAs solely based on primary deep sequencing data by considering the position and length-dependent patterns of reads aligned to short intervals on the genome. ALPS similarity rewards matching positions of reads of similar length in the two intervals. It can be computed efficiently and can be used to classify intervals of unknown function in various ways, one of which we have presented here.

ALPS only considers data that is available by a deep sequencing experiment and makes no further assumption about the common secondary structure of an ncRNA class. Such a scoring system is important not only because the RNA secondary structure prediction is not always accurate, but also because some ncRNA classes may not even have a common secondary structure. As long as members of a class share a similar pattern of read lengths and positions, our method is able to detect it. For instance, there is no method available to accurately detect usRNAs (Li *et al.*, 2009) in deep sequencing data in an automated manner. Since usRNAs are associated with RISC and their importance in post-transcriptional regulation has been shown, it is of great importance to provide a tool for their detection. Since they are characterized by their short length and fixed positions, ALPS similarity can be expected to identify them accurately in a deep sequencing experiment.

Our method can be used to support other, e.g. structure based, methods for the discovery of (specific) ncRNA classes by incorporating our similarity scores into the respective probabilistic model or machine learning scheme. As discussed, parameters of our scoring system can be finetuned in favor of any class of ncRNAs. In addition, if read lengths and positioning is also characteristic for subclasses, our scoring can be used to recover this hierarchy and for instance divide the class of miRNAs into the subclasses of $5'$ and $3'$ donors.

It has been suggested that miRNAs are modified after maturation (Morin *et al.*, 2008). These modifications are detectable in a deep sequencing experiment, and if they are specific for miRNAs, incorporating them into a scoring system should further boost the identification of miRNAs. This can easily be incorporated into the calculation of the similarity score by extending the column vectors and defining appropriate matrices M. Even structural information could be integrated the same way.

We have shown that only considering positions and lengths of deep sequencing reads already allows to accurately identify abundant miRNAs and tRNAs in a large-scale dataset. Our scoring system was not biased toward the identification of a specific class of ncRNAs and as a consequence, we expect it not only to be useful for the classification of known ncRNA types, but also for novel classes, as long as they exhibit a characteristic pattern of deep sequencing reads.

Funding: This work was supported by the German Bundesministerium für Bildung und Forschung (NGFN-Plus #01GS0801).

Conflict of Interest: none declared.

REFERENCES

Aravin,A.A. *et al*. (2001) Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. *Curr. Biol.*, **11**, 1017–1027.

Babiarz,J.E. *et al*. (2008) Mouse ES cells express endogenous shRNAs, siRNAs, and other microprocessor-independent, dicer-dependent small RNAs. *Genes Dev.*, **22**, 2773–2785.

Bachellerie,J.P. *et al*. (2002) The expanding snoRNA world. *Biochimie*, **84**, 775–790.

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Bentwich,I. (2005) Prediction and validation of microRNAs and their targets. *FEBS Lett.*, **579**, 5904–5910.

Chan,P.P. and Lowe,T.M. (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.*, **37**, D93–D97.

Czech,B. *et al*. (2008) An endogenous small interfering RNA pathway in Drosophila. *Nature*, **453**, 798–802.

Doshi,K.J. *et al*. (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, **5**, 105.

Dowell,R.D. and Eddy,S.R. (2004) Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, **5**, 71.

Fisher,S.R.A. (1970) *Statistical Methods for Research Workers*, 14th edn revised. Oliver & Boyd, Edinburgh.

Friedlander,M.R. *et al*. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.*, **26**, 407–415.

Gerard,M.A. *et al*. (2010) The scaRNA2 is produced by an independent transcription unit and its processing is directed by the encoding region. *Nucleic Acids Res.*, **38**, 370–381.

Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.

Griffiths-Jones,S. *et al*. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.

Grimson,A. *et al*. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.

Haussecker,D. *et al*. (2010) Human tRNA-derived small RNAs in the global regulation of RNA silencing. *RNA*, **16**, 673–695.

Higgs,P. and Morgan,S. (1995) Thermodynamics of RNA folding. when is an RNA molecule in equilibrium? In *Advances in Artificial Life*, Springer, Berlin/Heidelberg, pp. 852–861.

Hofacker,I.L. *et al*. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem. Chem. Mon.*, **125**, 167–188.

Hofacker,I.L. *et al*. (2004) Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, **20**, 186–190.

Kato,M. *et al*. (2009) Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during Caenorhabditis elegans development. *Genome Biol.*, **10**, R54.

Langmead,B. *et al*. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Li,Z. *et al*. (2009) Characterization of viral and human RNAs smaller than canonical MicroRNAs. *J. Virol.*, **83**, 12751–12758.

MacRae,I.J. *et al*. (2007). Structural determinants of RNA recognition and cleavage by Dicer. *Nat. Struct. Mol. Biol.*, **14**, 934–940.

Morin,R.D. *et al*. (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.*, **18**, 610–621.

Rathjen,T. *et al*. (2009) High throughput sequencing of microRNAs in chicken somites. *FEBS Lett.*, **583**, 1422–1426.

Shi,W. *et al*. (2009) A distinct class of small RNAs arises from pre-miRNA-proximal regions in a simple chordate. *Nat. Struct. Mol. Biol.*, **16**, 183–189.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Thompson,D.M. and Parker,R. (2009) Stressing out over tRNA cleavage. *Cell*, **138**, 215–219.

Will,S. *et al*. (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, **3**, e65.

Zien,A. *et al*. (2000) A simple iterative approach to parameter optimization. *J. Comput. Biol.*, **7**, 483–501.