Technical Report CMP-C07-02: Time Series Data Mining Algorithms for Identifying Short RNA in *Arabidopsis thaliana*

Anthony Bagnall Simon Moxon David Studholme

Abstract

The class of molecules called short RNAs (sRNAs) are known to play a key role in gene regulation. Th are typically sequences of nucleotides between 21-25 nucleotides in length. They are known to play a key role in gene regulation. The identification, clustering and classification of sRNA has recently become the focus of much research activity. The basic problem involves detecting regions of interest on the chromosome where the pattern of candidate matches is somehow unusual. Currently, there are no published algorithms for detecting regions of interest, and the unpublished methods that we are aware of involve bespoke rule based systems designed for a specific organism. Work in this very new field has understandably focused on the outcomes rather than the methods used to obtain the results. In this paper we propose two generic approaches that place the specific biological problem in the wider context of time series data mining problems. Both methods are based on treating the occurrences on a chromosome, or "hit count" data, as a time series, then running a sliding window along a chromosome and measuring unusualness. This formulation means we can treat finding unusual areas of candidate RNA activity as a variety of time series anomaly detection problem. The first set of approaches is model based. We specify a null hypothesis distribution for not being a sRNA, then estimate the p-values along the chromosome. The second approach is instance based. We identify some typical shapes from known sRNA, then use dynamic time warping and fourier transform based distance to measure how closely the candidate series matches. We demonstrate that these methods can find known sRNA on *Arabidopsis thaliana* chromosomes and illustrate the benefits of the added information provided by these algorithms.

1 Introduction

Short RNAs (sRNAs) are a class of sequences which are typically between 21-25 nucleotides in length and are known to play a key role in gene regulation [13, 20, 6, 18]. Short RNAs include both microRNAs (miRNAs), endogenous short-interfering RNAs (siRNAs) and trans-acting silencing RNAs (tasiRNAs). Both miRNAs and siRNAs are processed from a longer double-stranded RNA precursor by RNAseIII type enzymes of the Dicer family [9]. The double-stranded precursors are then unwound and the sRNA is recruited into the RNA induced silencing complex or RISC [8]. The sRNA then acts as a guide in the translational regulation or cleavage of target mRNA sequences by RISC.

Until recently, the detection and sequencing of sRNAs has been a laborious process, typically taking several months to obtain hundreds of candidate sequences. However, recent technological developments such as those at 454 Life Sciences [7] have enabled the high-throughput sequencing of sRNAs. This has made it possible to obtain hundredes of thousands of potential sRNAs from a single experimental sample. This breakthrough provides biologists with new opportunities to discover novel miRNA, tasiRNA and other species of RNA. It is widely assumed that the majority of the sRNA species present in a sample have originated by transcription of the genomic DNA sequence. Therefor an important and early step in the analysis workflow is identifying all occurrences of the observed sRNA sequences on the chromosomes. The next step is to identify *regions of interest* (ROI) where there are "unusual" levels of sRNA matching. Only a subset of the ROIs are likely to represent miRNA or tasiRNA loci. Validation of candidate sRNA loci in the molecular biology laboratory is a time-consuming and expensive procedure. It is clearly of upmost importance that the method for finding ROI prioritizes the most promising candidates without generating many false negatives. Section 2 provides a brief introduction into the biological importance of sRNA detection. The specific data sets used in our experiments are described in Section 3. At this early stage in the project we concentrate on the better understood problem of detecting of miRNA in the model plant *Arabidopsis thaliana*. There are over 100 known miRNA for *A. thaliana* and this provides a test data set for algorithm development. Ultimately we are more interested in the harder and less understood problem of detecting siRNAs, which exhibit a much more subtle hit count pattern. The method of generating the hit counts is formalised in Section 4, as is the generic region of interest problem.

Currently, there is no accepted best algorithm for finding ROI, and there has been very little published on the methodology of detecting sRNA-generated loci in a DNA sequence. This paper describes two alternative approaches for identifying the regions of interest based on statistical tests and time series data mining algorithms. Our guiding principles behind developing methods for this problem are to identify algorithms that:

1. can identify known miRNA on existing data;
2. produce a ranking of candidate ROI for further investigation;
3. are easily adaptable to new data sets; and
4. have a minimum number of parameters.

We adopt a definition of unusual (or surprising) similar to that given in [16] in that we consider a time series unusual if it is unlikely to have been observed from some pre-specified distribution (either model based or data driven) assumed to represent some normal pattern of hit counts over the chromosome.

The algorithms we develop come from two perspectives, but both use a sliding window along the hit count series and measure unusualness on each window. This generates a continuous measure of deviation that can easily be used to derive a ranked list of potential sRNA. The first approach is to phrase the problem as an hypothesis test. The null hypothesis specifies the type of distribution we would expect to see over the sliding window if the region did not contain an sRNA generated locu/region. We can then evaluate a p-value. We evaluate two approaches to the null distribution and resulting test. The first approach assumes observations within a window are independent and is based on a modified Chi-test. The second test allows for dependence within the window by using a spectrum method described in [5]. The statistical algorithms are defined in Section 5.

The second approach is a non-parameteric, instance based method that essentially matches candidates windows against known patterns of variation (modelled on detected sRNA) using both a dynamic time warping [14] and Fourier based distance metric [12]. These methods are described in detail in Section 5. Section 6 compares the output from all approaches against ground truth and discusses the relative merits of each. Finally, Section 7 assesses the biological importance of the data mining approaches and the implications on future work.

2 Background and Related Work

DNA stores the genetic information in all living organisms. This information is transcribed into messenger RNA molecules, which are similar to DNA but contain only the information for a single gene in eukaryotic cells. This information is translated into proteins, which can be catalysts of biochemical reactions or form structural components of the cell. Some RNA, however, is not translated into proteins but is fed into a different pathway. This is known as ‘RNA silencing’. For this to occur the RNA molecule must form double-stranded regions, which are recognized by an enzyme called Dicer and cleaved into small pieces. The resulting sRNAs typically are 21-25 bases long and function in guiding the so-called ‘RNA-induced silencing complex’ (RISC) to any RNA that exhibits sufficient sequence similarity. By binding to the target RNA (usually protein encoded mRNA), the latter is marked out for degradation. Thus, RNA silencing is an important mechanism of gene regulation that is also employed in defence against pathogens and genome maintenance. MiRNAs are a class of sRNAs involved in regulating other genes. They are produced from long precursor RNA molecules that form typical ‘hairpin’ structures by establishing base-pairs between different parts of the same molecule. This leads to the formation of short double-stranded regions that are processed to yield the mature miRNA [13].

To date, approximately 100 distinct species of miRNAs have been identified in the model plant *A. thaliana* by using classical molecular biological methods (e.g. [23]). Additional miRNAs have been proposed on the basis of computational predictions. However, recent technological advances in molecular biology have led to new large datasets that should allow us to more completely describe the repertoire of small RNA molecules in a cell. [2] have developed a biochemical method for partially purifying sRNA molecules that are physically associated with ARGONAUTE1 (AGO1) in plant material. Since the AGO1 protein specifically binds to miRNAs and tasiRNAs, then the population of sRNA sequences isolated by this method are highly enriched for miRNAs and tasiRNAs rather than other classes of sRNA. Using this technique, combined with high-throughput sequencing technology [7] we have isolated and sequenced miRNA candidates from the model plant thale cress *A. thaliana*. This procedure generated approximately 35,000 sequence reads yielding a sample of over 78,000. We assume that the majority of these sequences were encoded in the genome of *A. thaliana*; by simple exact string matching we are able to map RNA sequences onto the genomic DNA sequence. The challenge then is to characterise/descibe the pattern of hits on the genomic sequence and hence infer biological insights, including identification of novel miRNAs.

3 The *Arabidopsis thaliana* Dataset

The original query set of sRNA contains 78,208 sequences of between 17 and 26 base pairs in length. This set is used to discover regions of interest on 5 chromosomes which contain: 30,432,380; 19,690,779; 23,451,891; 18,584,949; and 26,967,032 base-pairs re-

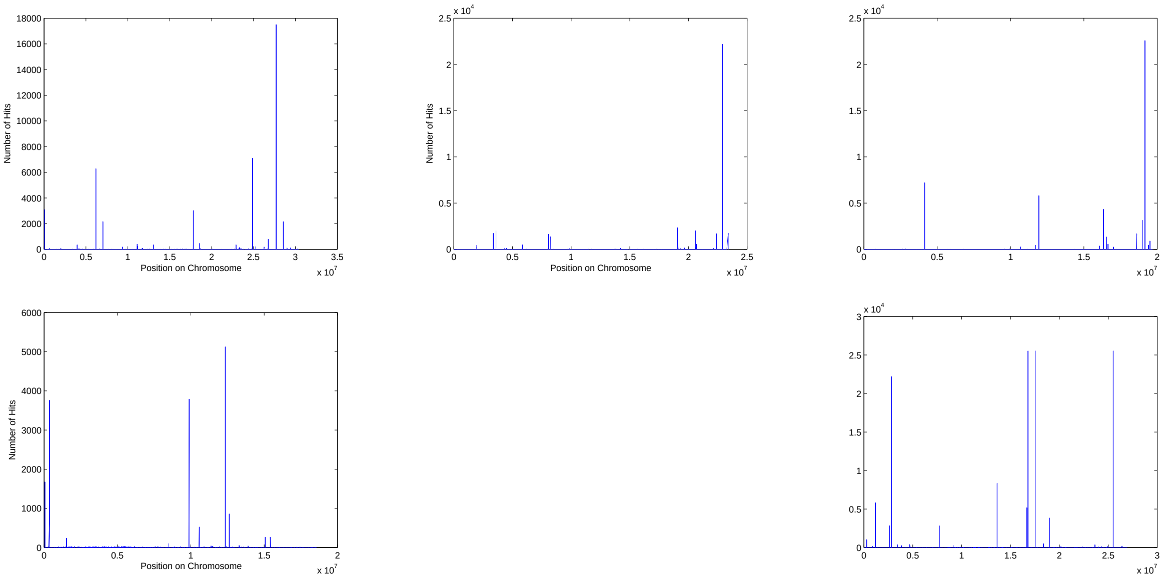


Figure 1: Hit counts for *A. thaliana*

spectively. The hit count series are very sparse, with: 253,100; 205,026; 274,942; 215,647; and 298,541 non-zero locations on each chromosome respectively. Figure 1 shows the hit count series for chromosomes 1 to 5.

Clearly, there are massive peaks in the hit count. These correspond to known miRNA and be can detected trivially. The real challenge is to detect regions that may be significant even though the hit pattern is much less pronounced and to distinguish between regions where the matches are probably just the result of chance, and those where there is a distinct pattern. To demonstrate the more subtle problem, Figure 2 shows the hit count for chromosome 1 at various resolutions. Figure 2 shows the region between positions 6,000,000 and 20,000,000. There are 9 known miRNA on this region. Two are obvious, with thousands of matches at the peak values. Below this are five smaller peaks (two very close together), which are also trivial to detect. Figure 2(b) shows the hit count for the first 2000 base pairs of the region after the large peak shown in Figure 2(a). There is a single known RNA in this area, between 18,000,000 and 18,050,000. There is a visibly different pattern of hits to the surrounding area, but it is harder to distinguish between the miRNA region and the surrounding area which may just be noise. Figure 2(c) shows a randomly selected region on chromosome 1, containing no known sRNA.

The known miRNA differ considerably in scale. The maximum number of hits within any miRNA for our *A. thaliana* sample ranges from 1 to 25517. However, there is a large degree in uniformity in the shape of the hit count distribution. Figure 3 shows three typical patterns of hit count within the known RNA. The series range from the more obvious unimodal pattern to more complex city block type patterns. It is thought that siRNA will exhibit an even more complex pattern of hits.

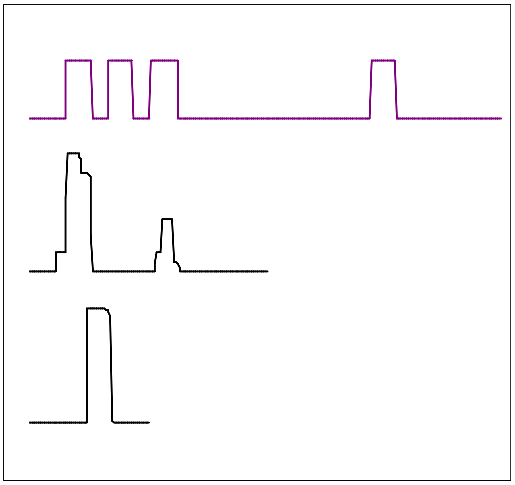
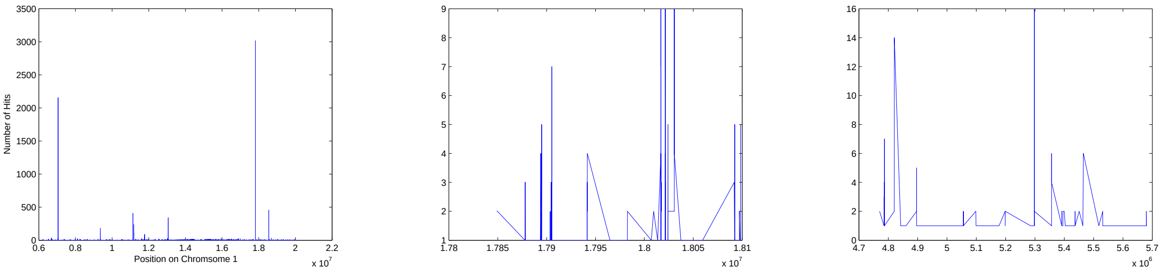


Figure 3: Three typical miRNA series.

4 Generalised Problem

In order to clarify the problem to the non-specialist and to motivate some of the approaches adopted we formally describe how we will generate the hit count series using



(a) (b) (c)

Figure 2: Hit counts for chromosome 1 at locations and resolutions. The region shown in figure (a) contains 9 known sRNA, the region in figure (b) 1 known sRNA and the region in figure (c) no known sRNA

a set of candidate sRNA and the chromosome DNA. We then define the problem of detecting regions of interest.

4.1 Generating the Hit Counts.

We are given a candidate set of Q of potential sRNA and a chromosome c for the organism in question. Q is a set of m vectors, $Q = \{\vec{q}_1, \vec{q}_2, \ldots, \vec{q}_m\}$ and $\vec{c}$ is a vector of nucleotides length n. Both $\vec{q}_i$ and $\vec{c}$ are sequences of letters from the alphabet A, C, G, T.

The length of any $\vec{q}_i$ is denoted m_i. The problem is to find and rank the subsequences of $\vec{c}$, referred to in this paper as regions of interest or ROI, in which the elements of Q are unusually represented for some definition of unusual (see Section 5). We denote a subsequence of $\vec{c}$ from j to k as $\vec{c}_{j,k}$. A single element of c is denoted c_j. Two sequences are said to be matched if and only if they are the same length and each element matches. We define a function to match two sequences as follows:

```
isMatched($\vec{q}, \vec{p}$) return boolean
    let $m \leftarrow$ size($\vec{q}$), $n \leftarrow$ size($\vec{p}$)
    if    $m = n$ and $q_i = p_i \ \forall \ i = 1 \ldots m$,
          then return true
    else return false
```

At any point c_j, a candidate $\vec{q}$ is said to be matched if isMatched is true for any substring of $\vec{c}$ that contains c_j. More formally, we can overload isMatched as follows

```
isMatched ($\vec{q}, \vec{c}, j$) return boolean
    let $m \leftarrow$ size($\vec{q}$)
    if isMatched($\vec{q}, \vec{c}_{i,k}$) for any $i, k \in [j - m \ldots j + m]$
          then return true
    else return false
```

The first step in finding the ROI is to derive the *hit count sequence*, $\vec{h}$. The hit count sequence records how many of the candidate set Q are currently matched at any position. Thus for position j, the hit count of set Q on chromosome c is simply

$$h_j = \sum_{\vec{q} \in Q} isMatched(\vec{q}, \vec{c}, j).$$

However, the hit count calculation is complicated by the fact that chromosomes are double stranded, and sRNA can match on either strand. It may be desirable to treat up and down hits differently. However, for the purposes of this research we combine the hit count of the up and down strands. The down strand element of a base pair is simply the complement of the upstrand, where the complement function maps the nucleotides $A \rightarrow T$, $C \rightarrow G$, $G \rightarrow C$ and $T \rightarrow A$. In addition, matching occurs in the opposite direction on the down strand. Rather than create a second string for the down strand, it is more computationally efficient to invert the and complement the candidates in Q using the operation inverseComplement

```
inverseComplement($\vec{q}$) return $\vec{p}$
    let $m \leftarrow$ size($\vec{q}$)
    for $i \leftarrow 1$ to $m$
        $p_i \leftarrow$ complement($q_{m+1-i}$)
    else return $\vec{p}$
```

The final hit count series $\vec{h^*}$ is found with the query set enhanced by the complement inverse. So if $Q' = q'_1, ..., q'_n$ where $q'_i =$inverseComplement(q_i), then the enhanced query set is $Q^* = Q \bigcup Q'$. The hit count series is then generated from the enhanced query set,

$$h_j = \sum_{\vec{q} \in Q^*} isMatched(\vec{q}, \vec{c}, j).$$

We illustrate generating a hit count series with an example. Suppose our query set is

Q={$\vec{q_1}$=ACGT, $\vec{q_2}$=ACG, $\vec{q_3}$=GT, $\vec{q_4}$=CT, $\vec{q_5}$=ACGTCA}

and the chromosome is

$\vec{c}$=(GGGAAAACGTACGGGCATTTTAACGTCA)

The hit count for the up strand is then

$\vec{h}$=00000022211110000001100333211000.

The complement inverse set, Q' is

Q={$\vec{q_1}$=ACGT, $\vec{q_2}$=CGT, $\vec{q_3}$=AC, $\vec{q_4}$=GA, $\vec{q_5}$=TGACGT}

Note that ACGT appears in both query sets. The inverse hit count is

$\vec{h}'$=001100232211000000000000121100000

and the hit count we would actually use is a combination of $\vec{h}$ and $\vec{h}'$
$\vec{h}^*$= 00110045432210000001100454311000

Our objective then, is to find regions of interest on the chromosome by consideration of just the hit count series. We may, for example, specify the regions (7,13) and (24,29) with sequences ACGTACG (hit counts 454322) and ACGTCA (hit counts 454311) as the interesting regions, but decide that (3,4) and (20,21) are in fact just noise. Note that the real regions of interest can exhibit some subtle variations in hit count pattern.

4.2 Detecting regions of interest with sliding windows.

The approaches we adopt involve running a sliding window across the hit count data, then using some measure of unusualness to form a second series measuring deviation from some expected normal behaviour. We call this the *significance series*, as it measures the moving p-value for an hypothesis test with a null *window does not contain a sRNA*. Figure 4 illustrates this process.

The alternative algorithms for generating a significance series are described in Section 5. In the rest of this Section we describe the three experimental procedures that are independent of the method used to form the series: setting the window size; defining the context of what is meant by unusual; and forming the ROI from the significance series.

The window width determines the granularity of the ROI created from the significance series. We use a simple non-parameteric test to determine the window width. Our starting premise is that, on the one hand, we do not want to classify regions with just a single hit as unusual, and on the other hand we want to keep the window width as small as possible in order to detect small variations in pattern. Thus we set the window size to be the smallest value where single hit windows constitute less than 5% of the windows with at least one hit. We find the value through enumerating all windows with at least one hit. Since we are using Fourier transforms, we then round up the window size to the nearest power of 2. For these experiments, we found the window size that meets our criteria was 512.

The model based algorithms described in Section 5 estimate the probability of any particular window having been observed when an sRNA is not present. This presupposes some definition of "normal" behaviour when sRNA are not present. This problem can be considered a time series anomaly detection problem [4, 15] (also called the detection of novelties [22], faults and temporal change). A key problem in any anomaly detection algorithm is to define what is meant by normal (or, conversely, what is meant by abnormal). There are three basic approaches to this problem adopted in the literature. The first involves determining the baseline by comparing the current window to the periods directly preceding it. The comparison can either be directly with the (possibly transformed) data [17, 15] or involve a construction of a model [25, 22]. This approach is not suitable for our problem, since sRNA may appear close together on the chromosome. The second method is to look for the windows most dissimilar to all other windows in the given data set. For example, Keogh, Lin and Fu [3] define *discords* as the most dissimilar non self match, then find discords by discretising the series using SAX [21] then embedding the global distribution in an augmented trie data structure. Whilst this is potentially useful for the sRNA ROI problem, and adapting SAX to use a long tailed Poisson distribution may yield useful results, we do not use it at this time. This is principally because of the length of the chromosomes and the large skew in the distribution of hits (98% of base pairs have no matches in the candidate set). Instead, we adopt the third technique to defining normality, which is to have the user specify either a typical data set [14] or typical pattern of variation [21].

Our approach is two-fold: the first is a model based. We estimate a distribution for normality based on the observations of hit counts for our candidate set on the entire chromosome.

The second approach is instance based (essentially a time series query by content problem [15]). We take an example known sRNA and measure the distance between the current window and the known sRNA. This gives us a continuous series of distances, from which we can extract ROI.

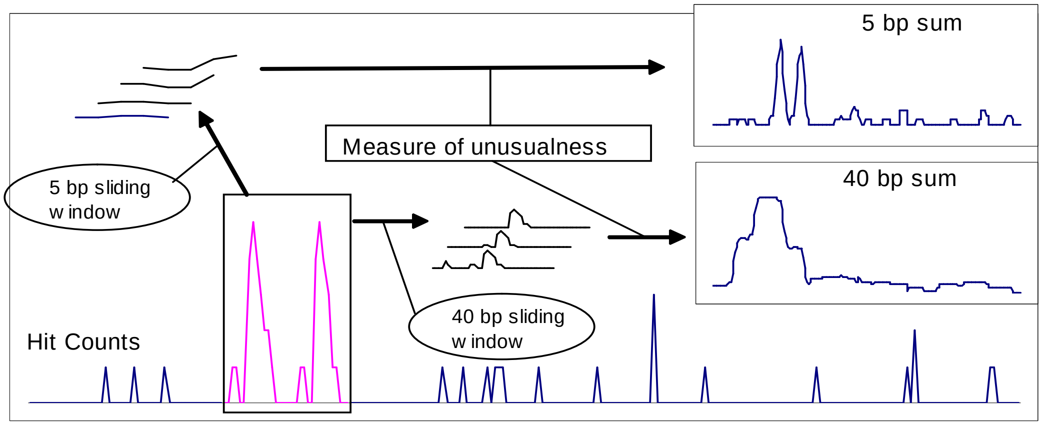


Figure 4: Example of the procedure for detecting unusual features in the hit count series. A sliding window creates subseries which are then evaluated with some unusualness measure. This creates a second series, the significance series, that can be used to determine ROI. In this example, the unusualness measure is simply the sum of the number of hits in the window.

Once the significance series has been found, forming the ROI for model based methods involves identifying all contiguous intervals where the p-value is above a significance threshold, set to 1% for our experiments. For the instance based approach, we can either use an traditional query by content procedure and report the k nearest sliding windows, or we can use an heuristic to generate ROI shorter than the window length from the continuous distance metric.

5 Algorithms for Estimating Significance

The basic estimation problem is as follows: given a window length w of hit count observations (512 observations in our case), what is the probability of this window containing an sRNA? The secondary problem is then to find the location of the sRNA within the window. The principle problem in developing algorithms for finding ROI is that the subject is so new there is very little training data available. The distinctive pattern of variation characterising sRNA is not well understood, hence the analysis is by definition highly exploratory. This is the primary reason we adopt four different methods to estimate probability of a region containing sRNA. As the body of knowledge increases, we can further test and develop the alternative approaches.

5.1 Model Based 1: Histogram Method

The first approach we adopted treats the hit count data for any window as an independent sample. Given a sample of observations of w random variables X_i, we can form a histogram of observed occurrences and compare it to an expected histogram with a χ^2 test. We can estimate the expected histogram from the entire chromosome. However, the large number of zero observations leads to very small expected values. Traditionally, cell values with an expected value of less than 5 are merged, so that the statistic is not unduly skewed by a single large observation. With a window size of 512, this leaves just two cells, 0 hits and 1 or more. To overcome this problem, we consider the expected values of hits for series given there is at least one hit. This gives us a histogram of six cells with expected values $\vec{e} = (e_0, \ldots, e_5)$, where e_0 represents the expected number of zero hits and e_5 the expected number of 5 or more hits. For any window with observed frequencies $\vec{b} = (b_0, \ldots, b_5)$, the distance is

$$d(\vec{b}, \vec{e}) = \frac{\sum_{i=0}^{5}(b_i - e_i)^2}{e_i}$$

and the significance value is found from the χ^2 distribution with 5 degrees of freedom.

Curtailing the histogram in this way discards potentially useful information. We could have adopted more complex measures to overcome the problem of small expected values, such as Fisher's exact test [1]. However, its clear from the example sRNA shown in Figure 3 that the basic assumption of independence is incorrect. Hence we have adapted a spectrum based method to be used in this context.

5.2 Model Based 2: Fourier Variance Test

The histogram method is able to detect the most obvious sRNA with large hit counts. However, it will not be able to detect sRNA defined by a small number of hits next to each other. To detect this kind of unusualness, we need a measure that can detect autocorrelations within the window in addition to absolute level over the whole window. One approach is to use methods based on the Fourier transform. For a real valued time series $\vec{y}$, defined over discrete intervals $y_t, t = 1, \cdots, N$, the fourier transform represents y as a linear combination of sinusoidal functions with amplitudes p, q and phase w,

$$y_t = \sum_{k=1}^{n} \left(p_k \cos(2\pi w_k t) + q_k \sin(2\pi w_k t)\right).$$

If series $\vec{y}$ has fourier coefficients (p_i, q_i) then the periodogram of $\vec{y}$, denoted here $\vec{f}$, is the series $f_i = p_i^2 + q_i^2$. We can derive a test from the periodogram to detect similarity between elements of $\vec{y}$. Suppose σ^2 is the variance of random variables $Y_1, Y_2, \ldots$ from which we assume $\vec{y}$ is an observation. If $\vec{y}$ is a stationary series, Y_t can be written as

$$Y_t = \varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \ldots$$

where ε_t is a white noise sequence with constant mean and variance σ^2. If the random variables Y_t are independent then $b_j = 0$ for all j. Hence the sample variance s_p^2 will be a good estimator of the variance σ^2. Kolomogorov [19] also proved that

$$log(\sigma^2) = \int_{-\pi}^{\pi} log 2\pi f_t dt.$$

Following the work of Davis and Jones [5], we can use these estimators to formulate a test which compares $var(X_t)$ and σ^2 estimated from the spectrum. The test compares the null hypothesis of

$$H_0 : b_j = 0, j = 1, 2, \ldots$$

against the alternative

$$H_1 : b_j \neq 0, \text{for at least one } j = 1, 2, \ldots$$

We estimate σ^2 from the spectrum by

$$log(\hat{\sigma}^2) = \frac{1}{n} \sum_{t=1}^{n} \int_{-\pi}^{\pi} log 2\pi f_t dt.$$

where $n = N/2$ and $c_1 = log 2 + \psi'(1)$. $\psi'(z)$ is the derivative of the log gamma function,

$$\psi'(z) = \frac{1}{z} + \frac{1}{2z^2} + \frac{1}{6z^3} + O(\frac{1}{z^4}).$$

Using asymptotic results we can derive a test statistic

$$\frac{\log s_p^2 - \log \hat{\sigma}^2}{\sqrt{\frac{\psi'(1)}{n} - \psi'((N-1)/2)}}$$

whose distribution tends to the standard normal N(0,1). For more details see [10].

5.3 Instance Based 1: Dynamic Time Warping Distance

Defining the null hypothesis for normal behaviour in such a new field is hard. New sRNA are being regularly discovered, and there is no theoretical model for the shape new sRNA will exhibit in a hit count series. An alternative approach is to use previously proposed sRNA as example instances, then to look for the regions that are in some way most similar to previously identified sRNA. This instance based approach is possibly the most promising for the future. It requires a measure of similarity between our set of *exemplars* (previously identified sRNA) and the sliding window candidates. This is essentially a time series data mining query by content problem [15].

For similarity in shape, Dynamic Time Warping (DTW) is commonly used to mitigate against distortions in the time axis [15, 24]. Suppose an exemplar sRNA is denoted $\vec{q}$ and the candidate from a sliding window $\vec{c}$ and, without loss of generality, both series are of length n. If $M(\vec{q}, \vec{c})$ is the $n \times n$ pointwise distance matrix between $\vec{q}$ and $\vec{c}$, where $M_{i,j}$ is the squared difference between q_i and c_i. A warping path $W =< (a_1, b_1), (a_2, b_2), \ldots, (a_k, b_k) >$ is a set of points that define a traversal of matrix M. A valid warping path must satisfy the conditions $(a_1, b_1) = (1, 1)$ and $(a_k, b_k) = (n, n)$ and that $0 \leq a_{k+1} - a_k \leq 1$ and $0 \leq b_k - b_{k+1} \leq 1$ for all $k < n$. The DTW distance between series is the path through M that minimizes the total distance, subject to constraints on the amount of warping allowed. Let $\mathcal{W}$ be the space of all feasible paths and $w_j = M(q_{a_j}, c_{b_j})$ be the distance between element a_j of $\vec{q}$ and b_j of $\vec{c}$ for the j^{th} pair of points in warping path W. The distance for any path X is

$$D_X(\vec{q}, \vec{c}) = \sum_{i=1}^{k} x_i.$$

The DTW path W is the path that has the minimum distance, i.e.

$$W = \min_{X \in \mathcal{W}} (D_X(\vec{q}, \vec{c})),$$

and hence the DTW distance between series is

$$D_W(\vec{q}, \vec{c}) = \sum_{i=1}^{k} w_i,$$

The amount of warping allowed is determined by the warping window width parameter r, which places a constraint on the maximum difference between the warping indexes a_k and b_k.

5.4 Instance Based 2: Spectrum Likelihood Ratio Test

If series y has fourier coefficients (p_i, q_i) then the periodogram of y is the sequence $a_i = p_i^2 + q_i^2$. One benefit of using the observed periodogram values is that, if the data is stationary, we can deduce the distribution of each term. It can be shown that each a_i can be thought of as an observation of an independent random variable A_i with exponential density

$$g(a) = \frac{1}{2\alpha_i} \exp\left(-\frac{a}{2\alpha_i}\right) \qquad i = 2, 3, \cdots, n-1.$$

Since we have independence, the likelihood of our series is

$$L(a) = \prod_{i=1}^{n-1} \frac{1}{2\alpha_i} \exp\left(-\frac{a_i}{2\alpha_i}\right)$$

and the log-likelihood is

$$\ell(a) = \sum_{i=1}^{n-1} \frac{a_i}{2\alpha_i} \log\left(2\alpha_i\right).$$

See [11] for a more complete description of the statistical properties of the periodogram terms. We can use the likelihood function to determine the similarity of two series by constructing a hypothesis test. Assume for simplicity that the two series are the same length and have periodograms a_i and b_i. The hypothesis of equivalence between the series is just the hypothesis that the random variables of which the periodograms are an observation, A_i and B_i, have the same distribution for all i. The likelihood ratio test would be based on the ratio

$$\lambda = \frac{\prod_{i=1}^{n-1} \frac{1}{2\hat{\alpha}_i} \exp\left(-\frac{a_i}{2\hat{\alpha}_i}\right) \prod_{i=1}^{n-1} \frac{1}{2\hat{\beta}_j} \exp\left(-\frac{b_j}{2\hat{\beta}_j}\right)}{\left\{\prod_{i=1}^{n-1} \frac{1}{2\tilde{\alpha}_i} \exp\left(-\frac{a_i}{2\tilde{\alpha}_i}\right) \prod_{i=1}^{n-1} \frac{1}{2\tilde{\beta}_i} \exp\left(-\frac{b_i}{2\tilde{\beta}_i}\right)\right\}}$$

where the "hat" and "tilde" denote the maximum likelihood estimates under the null hypothesis of equality and the alternative of inequality. It is straightforward to show that $\hat{\alpha}_i = \hat{\beta}_i = \frac{1}{2}(a_i + b_i)$, while $\tilde{\alpha}_i = a_i$ and $\tilde{\beta}_i = b_i$. Hence we can show that

$$\lambda = \prod_{i=1}^{n-1} \left(\frac{2a_i}{a_i + b_i}\right) \left(\frac{2b_i}{a_i + b_i}\right)$$

and the likelihood ratio statistic is then

$$\Lambda = -2 \log \lambda = 2 \sum_{i=1}^{n-1} \left\{2 \log(a_i + b_i) - \log a_i - \log b_i\right\}$$

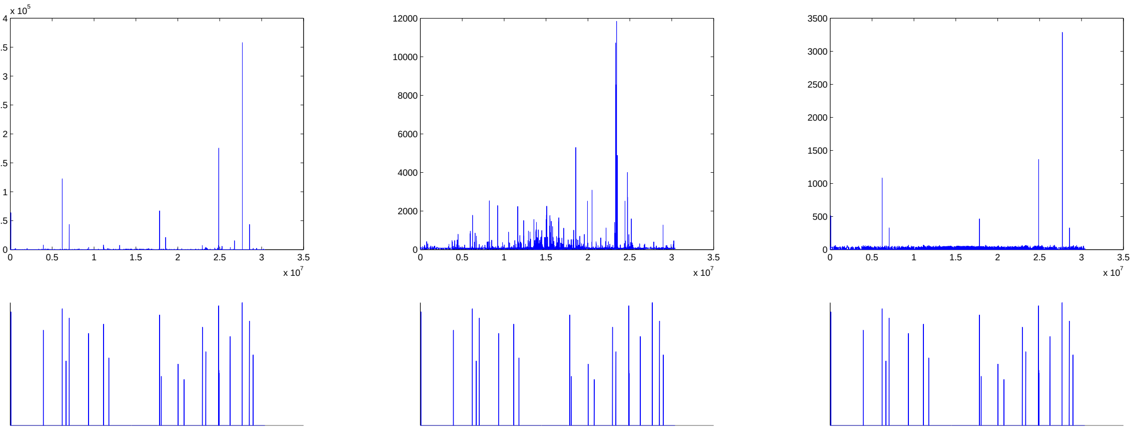
6 Results

We created significance series for the model based and the instance based methods described in Section 5, then generated regions of interest with the method outlined in Section 6. Our basic method for evaluating algorithms is to compare how many of the known miRNA are detected. However, it is worth noting several features of the evaluation before presenting results. Firstly, we have to look at the number of sRNA detected in relation to the number of regions proposed and the order they appear in the ranked list. We could easily maximize the number of miRNA detected by specifying all regions with a hit as ROI, but this would not be of much use. Secondly, the cut off significance level is somewhat arbitrary for the model based techniques (set to 1%) and needs to be empirically estimated for the instance based methods. This may influence the number of miRNA detected, as the lower we set alpha, the fewer ROI we detect.

We use a test derived from the sliding window sum statistic (as shown in Figure 4) as a base line comparison method. The test statistics for the significance series for the sum, chi-squared histogram test and the Fourier variance test for chromosome 1 are shown in Figure 5. Figure 5(a) shows the running sum series, Figure 5(b) the statistic for the histogram test and Figure 5(c) the statistic for the Fourier variance method. Beneath each figure is an indication of the location of the known miRNA. The length of line in the miRNA indicator graphs indicates the relative height of the peak hit count value.

We observe from these graphs that firstly, the sum test and the Fourier variance method both have a similar recognition pattern for the largest peaks, but the Fourier variance method seems to have higher variation at a lower level. The histogram method has a less clear cut pattern for the large peaks, and has a higher level of activity in the mid region where there are in fact no known miRNA. It is hard to interpret the importance of this: if could indicate that previously unknown miRNA are present, or that the algorithm is a poor discriminator for this kind of variation.

Table 1 describes the results for an increasingly strict significance level for inclusion into the ROI. It is worth noting that the miRNA used as a test set are easily detected by all the methods. However, we believe that using simple measures such as the sum of hits will be insufficient to detect all but the most obvious sRNA. We do not as yet have enough of the more complex sRNA to test this theory. However, Table 1 gives indications that the Fourier method may perform better in more testing conditions: As the definition for being classified as a ROI becomes stricter (i.e. we decrease



(a)

(b)

(c)

Figure 5: Significance series for: (a) sum of hits; (b) Chi-squared histogram test; and (c) Fourier variance test.

Table 1: Number of regions generated and percentage of miRNA found for a decreasing significance value cut off

Sum		Histogram		Fourier	
Nos	Correct	Nos	Correct	Nos	Correct
4263	100%	4109	100%	5659	100%
4010	100%	4308	100%	4406	100%
3040	100%	4659	100%	3709	100%
1997	98.97%	5131	100%	2595	98.97%
1022	94.85%	5497	100%	1958	96.91%
402.6	91.75%	6045	100%	1153	94.85%
296.8	89.69%	3964	69.07%	289.6	92.78%
40	73.20%	433	3.09%	51.2	88.66%

Table 2: miRNA identified by different algorithms on C1 and C4

Algorithm	miRNA found
Chromosome 1	
Sum	0,1,2,4-10,15,16,17,20-25
Histogram	1,10,12,17
Fourier	0-10,12,13,15-25
Chromosome 4	
Sum	0,1,2,5,6,8,9,10
Histogram	0,1,2,3,4,5,6,7,8,9,10
Fourier	0,1,2,5,6,7,8,9,10

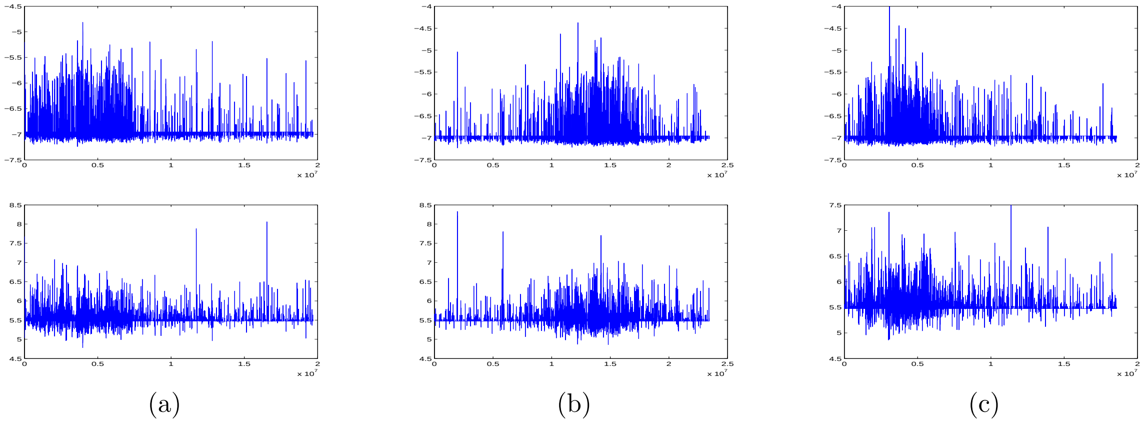
the significance threshold alpha), both Sum and Fourier generate less ROI, but Fourier seems to retain more of the known miRNA than Sum. Slightly surprisingly, as we decrease alpha, Histogram increases the number of ROI generated. This suggests than larger ROI are being split into smaller, neighbouring regions. This does not happen with Sum and Fourier, indicating different regions are being detected. Table 2 shows the known miRNA detected for chromosomes 1 and 4 under the lowest alpha for Sum and Fourier and lowest but one for Histogram used to generate the results shown in Table 1. It indicates that the techniques are discovering different miRNA. This suggests that using the methods in conjunction may yield better results.

Table 3 shows the proportion (over all 5 chromosomes) of ROI overlapping between the different methods for the highest level of alpha used to generate the results shown in Table 1. It indicates that the regions found by sum are generally also found by Fourier, but the converse is not always true. There is much less overlap between Histogram and the other algorithms. This is not unexpected. The fact that the Histogram algorithm truncates the observed frequencies means that much less prominence is given to very high peaks. Since high peak miRNA can be identified by eye, this is not necessarily a disadvantage. Histogram is more likely to find long flat regions with an unusual number of 1,2 and 3 hits.

Table 3: Degree of overlap between the ROI. The percentage is the proportion of ROI generated by the algorithm given in the row that are also generated by the algorithm in the column.

	Sum	Histogram	Fourier
Sum		67.20%	99.65%
Histogram	49.73%		45.82%
Fourier	81.70%	50.76%	

For the instance based methods, we use a single



(a) (b) (c)

Figure 6: Inverse distance series for DTW (top) and LR (bottom) for: (a) Chromosome 2; (b) Chromosome 3; and (c) Chromosome 4.

miRNA as an exmplar. The candidate was chosen because it has a slightly more complex shape than the average miRNA. Figure 6 shows the inverse of the log of the distance for both the DTW and Likelihood ratio method (we have transformed the graph so that large peaks indicate similarity to the query series). There is a similar pattern of variation, and some overlap between the generated regions. Figure 7 shows five example series detected by both algorithms that are strong candidates for being miRNA.

7 Conclusions

In this paper we have described the new bioinformatics problem of detecting sRNA from a very large sample of candidates and the occurrence of these candidates on the chromosome. We have phrased the problem as time series data mining query by content tasks and anomaly detection tasks. Four algorithms have been proposed: three are adaptations of previously used techniques, and the algorithm based on the Fourier variance test has, to our best knowledge, never been used in a time series data mining context before. We have applied these methods to data from the model plant *A. thaliana*. The data sets we have used will shortly be freely available for research purposes (contact the first author for details). There is a need to detect sRNA in all species, and interest in this problem is going to increase as more data becomes available. We have demonstrated that the model based algorithms algorithms are able to detect the known miRNA for *A. thaliana* and that they generate a diverse set of regions of interest which could be usefully combined to determine the most promising regions for investigation. Instance based methods demonstrate the potential for a more focused search for particular shapes in the hit count series.

The next step of this research will be to look at clustering and semi-supervised clustering of the ROI. Ideally, we would like to be able to classify ROI as indicative of different species of sRNA.

References

[1] Fisher R. A. The logic of inductive inference (with discussion). *J. Roy. Statist. Soc.*, (98), 1935.

[2] N. Baumberger D. C. Baulcombe. Arabidopsis ARGONAUTE1 is an RNA slicer that selectively recruits microRNAs and short interfering RNAs. In *Proc. Natl. Acad. Sci. U.S.A*, 2005.

[3] E. Keogh J. Lin A. Fu. HOT SAX: efficiently finding the most unusual time series subsequence. In *Proc. 5th IEEE International Conference on Data Mining*, 2005.

[4] D. Dasgupta and S. Forrest. Artificial immune systems in industrial applications. In *Proc. 2nd Intl. Conf. on Intelligent Processing and Manufacturing of Materials*, 1999.

[5] H. Davis and R. Jones. Estimation of the innovation variance of a stationary time series. *J. Am. Statistical Association*, 63:141–149, 1968.

[6] F. Vazquezc et al. Endogenous trans-acting sirnas regulate the accumulation of arabidopsis mrnas. *Molecular Cell*, 16:69–79, 2004.

[7] M. Margulies et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380, 2005.

[8] S.M. Hammond. An rna-directed nuclease mediates post-transcriptional gene silencing in drosophila cells. *Nature*, 404:293–296, 2000.

[9] G. Hutvagner. A cellular function for the rna-interference enzyme dicer in the maturation of the let-7 small temporal rna. *Science*, 293:834–838, 2001.

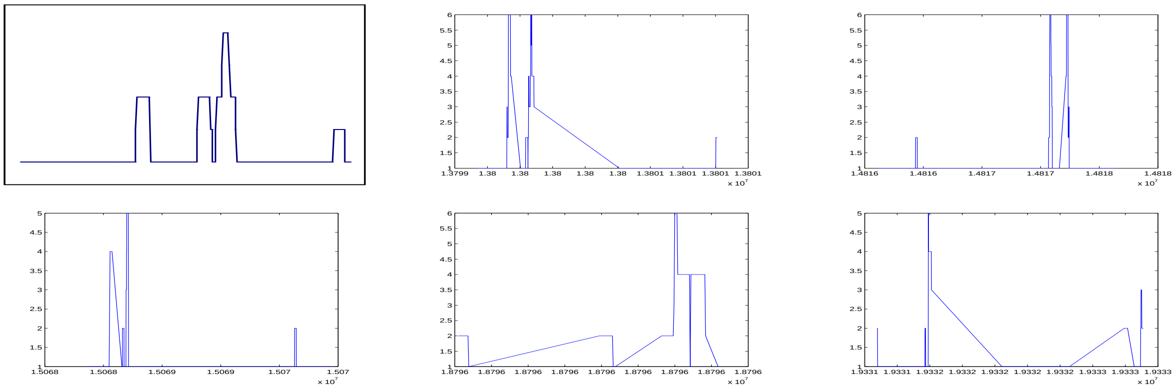


Figure 7: Examples of ROI found by both DTW and LR methods. The query miRNA is the top left graph.

[10] G. Janacek. Estimation of the minimum mean square error of prediction. *Biometrika*, 62(1), 1974.

[11] G. J. Janacek. *Practical Time Series.* Ellis Horwood, 2001.

[12] G. J. Janacek, A. J. Bagnall, and M. Powell. A likelihood ratio distance measure for the similarity between the fourier transform of time series. In *Proc. 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2005*, 2005.

[13] F. Meins Jr., A. Si-Ammour, and T. Blevins. Rna silencing systems and their relevance to plant development. *Annu. Rev. Cell. Dev. Biol.*, 21:297–318, 2005.

[14] E. Keogh. Exact indexing of dynamic time warping. In *Proc. 28th International Conference on Very Large Data Bases*, 2002.

[15] E. Keogh and S. Kasetty. On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7(4), 2003.

[16] E. Keogh, S. Lonardi, and B. Chiu. Finding surprising patterns in a time series database in linear time and space. In *Proc. of the Eighth ACM SIGKDD Knowledge Discovery and Data Mining*, 2002.

[17] E. Keogh and M. Pazzani. Scaling up dynamic time warping to massive datasets. In *Proc. 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'99)*, 2000.

[18] V.N. Kim. Small RNAs: classification, biogenesis, and function. *Molecules and Cells*, 19(1):1–15, 2005.

[19] A. Kolmogorov. Sur l'interpretation et extrapolation des suite stationnaries. *Comptes Rendus Ac. Sc.*, 208, 1939.

[20] N.C. Lau, L.P. Lim, E.G Weinstein, and D.P. Bartel. An abundant class of tiny rnas with probable regulatory roles in caenorhabditis elegans. *Science*, 294:797–799, 2001.

[21] J. Lin, E. Keogh, S. Lonardi, and B. Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11, 2003.

[22] J. Ma and S. Perkins. Online novelty detection on temporal sequences. In *Proc. of the Ninth ACM SIGKDD Knowledge Discovery and Data Mining*, 2003.

[23] B.C. Meyers, F. F. Souret, C. Lu, and P.J. Green. Sweating the small stuff: microrna discovery in plants. *Curr. Opin. Biotechnol.*, 17(2):139–146, 2006.

[24] C. A. Ratanamahatana and E. Keogh. Three myths about dynamic time warping data mining. In *Proc. SIAM International Conference on Data Mining (SDM '05)*, 2005.

[25] W-K. Wong, A. Moore, G. Cooper, and M. Wagner. Bayesian network anomaly pattern detection for disease outbreaks. In Tom Fawcett and Nina Mishra, editors, *Proc. of the Twentieth International Conference on Machine Learning*, 2003.