PAPER ANALYSIS

Bioinformatics Class - 05 avril 2020

Guillaume De Gani

Towards a piRNA prediction using multiple kernel fusion and support vector machine

Jocelyn Brayet[1,2], Farida Zehraoui[1], Laurence Jeanson-Leh[2], David Israeli[2] and Fariza Tahi[1,*]

[1]IBISC EA 4526, UEVE/Genopole, IBGBI, 23 bv. de France, 91000 Evry, France and [2]Genethon, 1, bis rue de l'Internationale, 91002 Evry Cedex, France

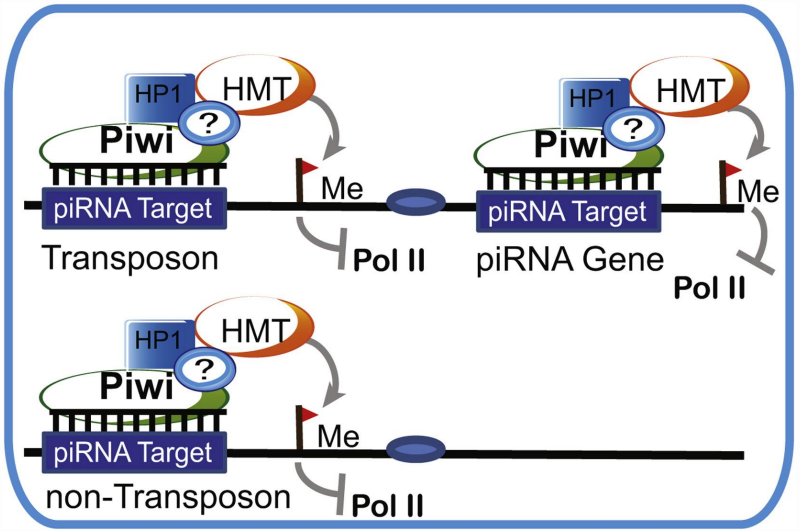
What is piRNA and why sequencing it in a effective matter is necessary ?

piRNA or Piwi-interacting RNA is the largest class of small non coding RNA in animal cells. Usually it is composed of only 21 to 31nt (nucleotide) unlike miRNA in the range of 21 to 24nt.

These piRNA are involved in epigenetic and post-transcriptional silencing of transposable elements and other spurious or repeat-derived transcripts, but can also be involved in the regulation of other genetic elements in germ line cells .[3][4][5]

As mentioned in the paper studied the recent discovery indicating their involvement in diseases such as cancer makes their identification a necessity.

Proposed piRNA structure, with the 3' end 2'-O-methylation. Source Wikipédia



Epigenetic mechanism guided by piRNAs. Source: Yale Stem Cell Center and Department of Cell Biology

The research context of piRPred development.

As aforementioned in the article studied the lack of conserved characteristics makes their identification by computational method difficult.

However some noticeable features seem to emerge from their study like:

- A uridine nucleotide at the 5' first position of the transcript (Le Thomas et al., 2014).
- piRNA seems to be encoded in clusters of 1 to >100kb nt long arrays. [Brenneckeet al., 2007; Lauet al., 2006]
- Their vicinity with the telomeric and centromeric region of the chromosome. [Brenneckeet al., 2007; Le Thomas et al., 2014]

Current computational Methods

Linear classification method:

This technique is probably the simplest in the field of machine learning, however it is impossible to find non linear boundaries with this method which can limit its effectiveness with complexe data like piRNA. (Zhanget al., 2011)

Clustering

This method is based on grouping the data if they are similar. In this case finding patterns and redundant expressions as in HCS clustering algorithm[49][50] for example.(Junget al., 2014; Rosenkranz and Zischler, 2012)

Novelty of the piRPred algorithm.

The piRPred algorithm uses the three characteristics described in the previous slide in combination with the usage of k-mer strings to identify motifs in piRNA sequences.(Zhang et al.)

This approach of using all four of these features together could lead to an improvement in the results obtained by existing algorithms.

The following part of the analyses will review the implementation of this new approach and whether or not it lead to better results.

Kernel fusion support vector machine algorithme

Support-vector machine is a supervised learning method that is used in numerous classification problems. It addition to performing linear classification efficiently you can use the “kernel trick” to classify in higher dimension this is the method used by piRPred.

The different kernels will be described in the next part of the presentation.

Novelty of the piRPred algorithm.

Kernel 1

This kernel represents K-mer string and uridine position. The first element of the vector represents the presence or uridine at the 5' or its absence. The 32 other elements are the frequencies of the k-mer described by Zhang.

Kernel 2

This kernel represents the distance of the sequence to the closest pericentromeric and subtelomeric region of the genome. Note if the sequence is located in one of these region the value will be +inf.

Kernel 3

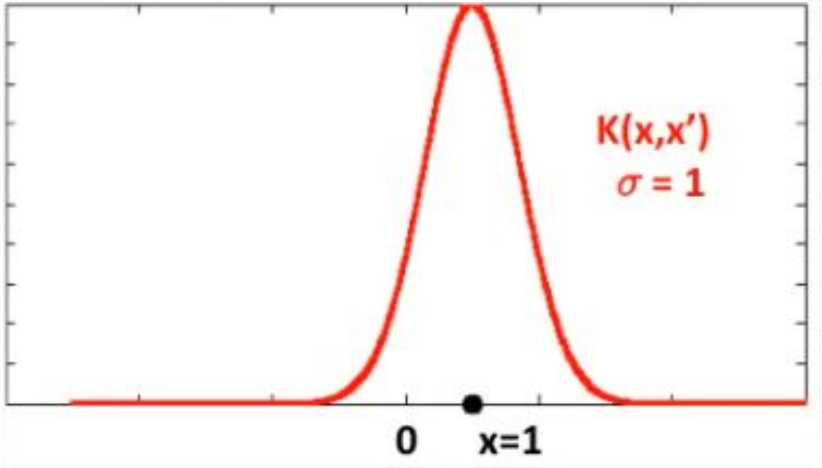
This kernel represents the distance ft the k-nearest neighbors of the sequence studied. The neighbors studied are within the training set and on the same chromosome as the targeted sequence.

Every kernel uses a specific Radial Basis Function which is the the following:

$$k(x,y) = exp(-\gamma||x-y||^2)$$

The value of γ is determined by the popular grid search method. Instead of brute forcing the best value for γ a heuristic approach is used which is less time consuming.

Gaussian kernel function function. With σ instead of γ

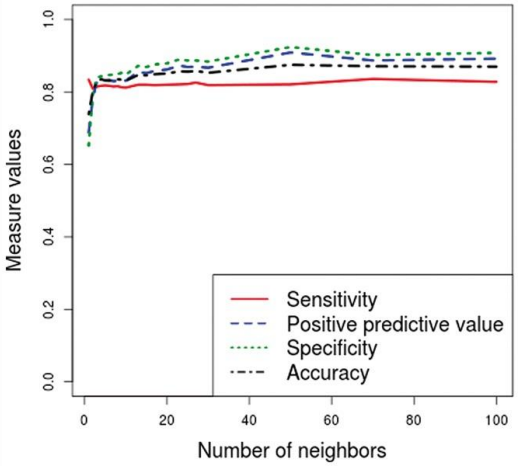


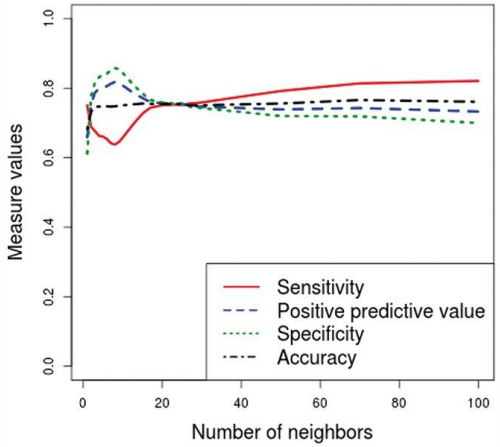
Selecting the number on neighbors studied in the third kernel.

The value of k is extremely important since it has a considerable impact on the calculation time of the programme. This is due to the data which is represented by a matrix of size (k+1)x(k+1) which leads to a complexity of $O(k^2)$.

So for computational purposes they choose the smallest value of k for which the accuracy was maximal. As we can see in the two figures accuracy was stable above 4 so this value was chosen.

However this parameter can still be tuned by the user which means that if a user has more computational power and or time he can run the algorithm with larger value of k to study the impact of the number of neighbors considered for sequence.





The importance of the Training dataset.

A machine learning project is only as good as its training set. Which is why in the paper they carefully designed a training set for the algorithm. Especially the Negative which needs to be close enough the positive cases to properly train against false positives.

Positive Data

This data comes from the piRNABank which contains 23 439 and 22 336 non redundant piRNA sequences from respectively Humans and Drosophila genomes.

Negative Data

The negative dataset was build with non-redundant sequences of several types:

- tRNA sequences of size 25 to 33 nt from the tRNA database
- Mature miRNA sequences from the miRBase
- Randomly chosen sequences of size 25 to 33 nt from exonic regions of protein-coding genes.

Paper Analysis

Cross Validation

To evaluate the the relevance of the different kernels used in the algorithm they are tested independently. Then they are combined using the mean method and the SPG-GMKL method. All this data is given in the table below.

Furthermore the evaluation is done using 5-fold-cross-validation which consists in separating the dataset in 5 equal partition. One partition is left out to be the controle and the algorithm is trained on the four remaining partitions. Then you average the 5 results obtained by leaving out a different partition each time

To evaluate the classification performance, we use several statistical measures: accuracy *ACC*, sensitivity *SE*, specificity *SP* and positive predictive value *PPV*. These measures are defined as follows:

- Accuracy $ACC = \frac{TP+TN}{TP+TN+FP+FN}$, which measures the percentage of samples that are correctly classified.
- Sensitivity $SE = \frac{TP}{TP+FN}$, which measures the accuracy on positive samples.
- Specificity $SP = \frac{TN}{TN+FP}$, which measures the accuracy on negative samples.
- Positive predictive value $PPV = \frac{TP}{TP+FP}$, which measures the percentage of correctly classified positive samples among all positive-classified ones.

where *TP*, *FP*, *TN* and *FN* are the numbers of true-positive, false-positive, true-negative and false-negative predictions, respectively.

Method	Human				Drosophila			
	ACC	SP	SE	PPV	ACC	SP	SE	PPV
Km	0.76 ± 0.03	0.75 ± 0.01	0.81 ± 0.01	0.75 ± 0.02	0.67 ± 0.01	0.70 ± 0.02	0.65 ± 0.01	0.66 ± 0.02
Kd	0.61 ± 0.02	0.55 ± 0.02	0.72 ± 0.03	0.59 ± 0.01	0.86 ± 0.02	0.88 ± 0.03	0.83 ± 0.01	0.86 ± 0.02
Kn	0.74 ± 0.01	0.82 ± 0.02	0.67 ± 0.03	0.80 ± 0.02	0.83 ± 0.03	0.82 ± 0.01	0.83 ± 0.04	0.82 ± 0.01
Km/Kd/Kn mean	0.81 ± 0.03	0.82 ± 0.02	0.78 ± 0.03	0.81 ± 0.02	0.87 ± 0.02	0.93 ± 0.01	0.81 ± 0.03	0.91 ± 0.02
Km/Kd/Kn SPG-GMKL	**0.86 ± 0.02**	**0.84 ± 0.01**	**0.88 ± 0.03**	**0.85 ± 0.02**	**0.89 ± 0.03**	**0.95 ± 0.02**	**0.83 ± 0.03**	**0.94 ± 0.03**
Zhang *et al.*	0.58 ± 0.05	0.82 ± 0.01	0.30 ± 0.04	0.63 ± 0.03	0.69 ± 0.02	0.92 ± 0.01	0.45 ± 0.02	0.85 ± 0.01

Note: ACC, accuracy; SP, specificity; SE, sensitivity; PPV, positive predictive value. In bold: The highest value in each column.

Conclusion

In conclusion this method seems to be more accurate than the previous algorithms. Furthermore this method can be improved for example if in the future a new characteristic of piRNA is discovered it it possible to simply add a kernel which could improve prediction results.

Or the training could be done on a larger dataset which could also greatly improve the performance of the programm.