



uOttawa

## University of Ottawa – Project UO3

**Title:** Novel algorithm for mapping miRNA targets

**Supervisor:** Ilya Ioshikhes

**Contact information:** [lioschik@uottawa.ca](mailto:lioschik@uottawa.ca)

**Field:** computational biology; bioinformatics

**Duration:** 6-8 weeks

### **Project context:**

It has been discovered that over 50% of human genes may be regulated, in part, by a posttranscriptional mechanism involving microRNAs (miRNAs). MiRNAs are small RNAs that regulate gene expression primarily through translational repression. Medical significance of the miRNAs is conveyed through regulation of various genes related to cancer, cardiovascular function etc. Precise knowledge of the miRNA binding sites (targets) is essential for understanding of specific mechanisms of the posttranscriptional regulation.

### **1. Existing algorithms for identifying miRNA targets.**

**1.a. *MiRanda*.** The *miRanda* algorithm consists of two basic steps, supplemented by statistical and phylogenetic estimations to identify potential targets. First a dynamic programming local alignment (essentially a modified Smith-Waterman alignment) is carried out between the query miRNA sequence and the potential genomic target sequence. The scores of the alignment are based on sequence complementarity and not on sequence identity. At the second stage of analysis, the Vienna package for RNA folding is employed to estimate the thermodynamic properties of a predicted duplex, in addition to those by alignment scores. MiRNAs conserved across several species get higher priority in the search.

**1.b. *TargetScan*.** This popular program combines thermodynamics-based modeling of miRNA/mRNA duplex interactions with comparative sequence analysis to predict miRNA targets conserved across multiple genomes. *TargetScan* searches the 3' UTRs for segments of perfect Watson-Crick complementarity to bases 2–7 of the miRNA numbered from its 5' end (so called “seed match”). The program then extends each seed match as far as possible and uses the RNA-fold program of the Vienna package to complete the alignment. The program produces the scores according to the sites' binding energy and searches for conserved regions in other species. It explicitly relies on targets conserved across species for its predictions.

## Stages d'été en recherche à l'international pour étudiants du premier cycle (SÉRI) Summer Undergraduate International Research Internships (SIRI)

---

**1.c. *Diana* – predicting human miRNA targets.** This program initially identifies putative miRNA/mRNA interactions based on binding energies between two RNAs paired imperfectly. A modified dynamic programming algorithm is applied calculating a pairing between the two sequences that yields the minimum free energy. The program uses two nucleotides at once to calculate the free energy between the two nucleotides of the miRNA paired with the two from the putative mRNA target.

Some sort of dynamic programming alignment is involved in all these programs, usually combined with a computationally time-consuming RNA folding procedure or other estimates of the RNA duplex binding energy. Remarkably, programs' emulations available online simply search for putative targets among pre-calculated data, instead of performing real interactive calculations. This makes it much harder to predict targets for novel miRNA sequences. Most of the programs search only for sites conserved across several species. Although this may lead to important discoveries, it surely misses the targets different between the species.

### **2. Biochemical principles of miRNA-target recognition.**

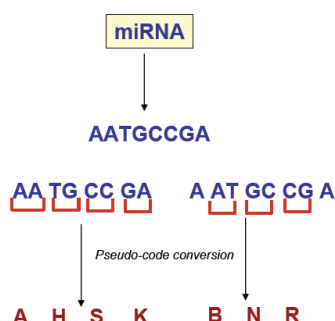
Although the biological importance of miRNAs has become quite clear, particular biochemical rules of recognition and regulation of target genes remain much less understood. One of such rules was formulated in the notion of the "seed match" which transformed during a time. However, Brennecke et al. systematically evaluated the minimal requirements for functional miRNA–target duplexes in vivo and distinguished classes of target sites with different functional properties. Target sites have been grouped into two major categories: 1. 5' dominant sites with sufficient complementarity to the miRNA 5' end that may function with little or no support from pairing to the miRNA 3' end, and 2. 3' compensatory sites that have insufficient 5' pairing and require strong 3' pairing for function. Examples and genome-wide statistical support were presented to show that both classes of sites are used in biologically relevant genes. That supports our position that all experimental constraints should be accounted in a single statistical model, and not added outside of the main framework.

### **Research Plan**

Currently existing algorithms for the miRNA target prediction typically use combination of miRNA/mRNA sequence alignment with energetic miRNA/mRNA binding considerations. A typical sequence alignment procedure uses a modified Smith-Waterman alignment based on evolutionary considerations. As mentioned by Smith and Waterman, the alignment is part of more general problem "to measure the minimum number of 'events' required to convert one sequence into another". That idea is well applicable in analysis of evolutionarily related sequences (genes, genomes and proteins), yet its straightforward application to the evolutionary unrelated sequences (like miRNA/mRNA) looks rather artificial. In any case it should be further combined with quite time-consuming RNA folding algorithm to bring into consideration energetic parameters of the miRNA/mRNA binding. In our project, we are suggesting development of a novel miRNA target search algorithm based on energetic rather than on evolutionary considerations. Although using basic concept of dynamic programming alignment by Smith-Waterman we modify it, using miRNA/mRNA binding parameters in the alignment itself, thus eliminating need in the most time consuming procedure – artificial RNA folding. It makes our program much faster than other algorithms (up to 100 times faster than miRanda, by our estimates).

### Step 1. Design of the modified miRNA/mRNA alignment procedure.

A typical Smith-Waterman algorithm for DNA/DNA alignment should be modified according to the rules of nucleotide complementarity for the miRNA/mRNA alignment. In that case G/C and A/T(U) matches are considered as perfect, unlike A/A, C/C, G/G and T/T in regular case. G/U wobble pair is usually also considered, with a score lower than for the perfect matches. We use RNA duplex stacking energies for alignment scores, to adjust the alignment to the biophysical nature of the problem. Binding energy of RNA/RNA duplexes depends on the dinucleotide composition thereof, not just on the mononucleotide one. Therefore we modify the sequence alignment implementing one working on dinucleotide matches and mismatches, not on those of mononucleotides. This is done by conversion of regular 4-letters nucleotide sequence to a 16-letters sequence (Fig. 1). Genomic sequences are converted accordingly. A Smith-Waterman-style alignment is then implemented for the converted sequences based on the free energy parameters for DNA and RNA folding.



**Figure 1.** Conversion of a regular mononucleotide sequence to dinucleotide sequence. Overlapping dinucleotide frames are shown separately for better view, yet they will coincide in the converted sequence (ABHNSRK).

Only the best alignments (scoring above certain cut-off) are reported by the program, with the sequences back-converted to the regular 4-letters alphabet. A preliminary version of the software was already compiled.

### Step 2. Initial evaluation of the performance of the algorithm and its optimization.

To evaluate the performance of the algorithm, we need to estimate its sensitivity (proportion of the true positive predictions TP out of the all true targets, recognized [TP] or missed [false negative, FN]) (Formula 1) and specificity (proportion of the TP predictions out of true and false predictions FP together) (Formula 2).

$$SN = TP / (TP + FN)$$

(Formula 1)

$$SP = TP / (TP + FP)$$

(Formula 2)

Presumably FP predictions are those done on the sequences without any biological meaning, e.g. shuffled sequences. We will shuffle the studied genomic sequences, and will map putative miRNA targets on the shuffled sequences. We will then optimize our algorithm in order to reduce the FP level but correctly identifying experimentally verified miRNA targets. Following parameters will be subject for optimization.

1) Scaling parameters. There are different levels of stringency for the miRNA-mRNA matches. It was accepted until recently that some miRNA nucleotides (primarily 2-7) should virtually perfectly match the mRNA sequence, whereas in the others mismatches are quite possible. That property explored by the Target Scan and latest miRanda version. In the latter, it is taken into account as a scaling factor, with scores in these positions multiplied by 2 versus the initial scores. The aforementioned Brennecke rules are basically

## Stages d'été en recherche à l'international pour étudiants du premier cycle (SÉRI) Summer Undergraduate International Research Internships (SIRI)

---

consistent with this notion for the 5' dominant sites. However the 3' dominant sites do not obey the rules mentioned here. In the framework of our algorithm, position-dependent scaling factors for dinucleotide sequence matches and gap penalties (see below) will be calculated upon the algorithm optimization.

2) Gap parameters. The gap parameters also may vary, both for the gap opening and extension. We will try scoring such gaps by their energetic parameters as well.

3) Mismatches scores. Smith-Waterman alignment for DNA usually gives a certain penalty for mismatches, as in miRanda program. We should optimize the score during development of our algorithm. In addition, energetic value of each mismatch is different, as it is also for different matches. We will assign specific scores for all the mismatches in the refined version of our alignment.

4) Conserved targets. MiRanda and other algorithms heavily explore conservation of putative miRNA targets across genomes of different species. We will explore that property as well in a consistent manner, but as an option only. That would allow users search for conserved or species-specific targets.

### **Required academic background**

Interests and basic knowledge in molecular biology, genetics, bioinformatics, and computing. Ability to work with scientific software and adjust it. Programming experience is a plus.