The Method Description of Target Gene Prediction

There are two main algorithms to predict target genes. They're described as follows:

1. The descriptions and computing processes:

MiRNA can combine with 3'-UTR regions of mRNA, performing post-transcription regulation. When combining, the second to eighth base from 5'-end of the miRNA take more important role, so the 7-length sequence is called the seed of a miRNA sequence [1]. According to the location of a seed combining with the 3'-UTR region, there're 3 conditions:

(1). 8mer-1a: the 7 bases are all paired. What's more, the base at 3'-UTR next to its 3' end is A;

(2). 7mer-8m: the 7 bases are all paired, but the base at 3'-UTR next to its 3' end is not A;

(3). 7mer-1a: similar to 8mer-1a, but the last base at the seed's 3' end is not paired;

The preceding types have decreasing intensity of combination. The intensity is measured by $score_{type}$. A miRNA sequence can have multiple potential combination locations at a 3'-UTR sequence, the combination type is one of the preceding three.

What's more, the following factors also influence the combination intensity:

(1). The position of the combination location [3]. The closer the location is to the end of 3'-UTR, the more intense. But the location is not allowed too close to stop codon.

(2). The density near the location of A or U. The denser the A/U bases, the more intense. Of course, the non-seed area of the miRNA must be also consider. The more bases are paired, the more intense is the combination.

(3). The seed-pairing stability (SPS). A smaller value (greater absolute value) indicates more intense combination.

(4). Target-site abundance (TA): the relative amount of the locations of a seed at many 3'-UTRs. A smaller TA indicates higher prediction efficiency.

Every factor is measured by a type of score, and the sum of all is called context+ score [4]. This score is used to evaluate the intensity of combination of a seed with a potential paired location at a 3'-UTR. A smaller value (greater absolute value) implies higher probability of a target gene.

The processes of computing context+ score:

$\mathrm{rawScore}_{\mathrm{pos}}$ means the minimum of the two distances from the combination location to the two end of 3'-UTR. Then the position score score_{pos} is computed by:

$$score_{\mathrm{pos}} = rep_{pos}\left(\frac{rawScore_{pos} - score_{posmin}}{score_{posmax} - score_{posmin}} - score_{posmean}\right)$$

In which rep_{pos}、$score_{posmin}$、$score_{posmax}$、$score_{posmean}$ are obtained from huge amount of samples. In addition, when the potential combination location is within 15 bases to a stop codon, this location is considered not to be a target site.

Given a potential combination location, retrieve a 30-length sequence at 3'-UTR upstream and downstream respectively. First, make an overall score and an AU score to be zero. For every base at the two sequences, if there are n bases from the seed to it, then add 1/n to the overall score. If the base is A/U, then add 1/n to the AU score too. $rawScore_{AU}$ is the quotient of the overall score divided by the AU score. As an addend of the context+ score, $score_{AU}$ is computed by:

$$score_{AU} = rep_{AU}\left(\frac{rawScore_{AU} - score_{AUmin}}{score_{AUmax} - score_{AUmin}} - score_{AUmean}\right)$$

In which rep_{AU}、$score_{AUmin}$、$score_{AUmax}$、$score_{AUmean}$ are obtained from huge amount of samples.

In addition to seeds, the 3'-end of a miRNA sequence is important with regard to target gene prediction, especially from the 13th to the 16th base beginning from 5'-end of the miRNA. Given a combination location according to the seed, consider a sub sequence of length 23 of the 3'-UTR, the seed being paired at its 3'-end. If a sub sequence of the miRNA, which is at its 3'-end, can be completely paired with a sub sequence of the 3'-UTR at 5'-end, then a score value can be obtained. Every preceding 13-16th base contributes 1 to the score and others 0.5 each. The maximum of this type of scores is called $rawScore_{3supp}$, which is a part of the context+ score:

$$score_{3\mathrm{supp}} = rep_{3supp}\left(\frac{rawScore_{3supp} - score_{3suppmin}}{score_{3suppmax} - score_{3suppmin}} - score_{3suppmean}\right)$$

In which rep_{3supp}、$score_{3suppmin}$、$score_{3suppmax}$、$score_{3suppmean}$ are obtained from huge amounts of samples.

The score of seed pair stability score_{SPS} and the score of target abundance are computed similarly. And seed's $score_{SPSmin}$、$score_{SPSmax}$、$score_{SPSmean}$、$score_{TAmin}$、$score_{TAmax}$、$score_{TAmean}$ and the raw seed pair stability score $rawScore_{SPS}$, and the raw target abundance score are relative values and normalized after considering all seeds. rep_{SPS}、rep_{TA} are estimated recession coefficients.

Lastly, the context+ score can be computed by:

$$score_{context+} = score_{pos} + score_{AU} + score_{3supp}$$

$$+score_{SPS} + score_{TA} + score_{type}$$

The smaller is this score value, the more likely is a target gene.

TargetScan also consider the conservation across species. When being conservative, the probability of conserved target (PCT) is computed. See the appendix.

2. The computer processes of miRanda

The method used by miRanda is based on dynamic programming (SW algorithm [5]) and computing free energy. When the score value of aligning sequences(miRNA and 3'-UTR) and the free energy are greater than corresponding pre-defined thresholds respectively, then it is considered that the gene, of which the 3'-UTR is a sub sequence, is the miRNA's target gene. During this process, the concrete alignment is obtained.

When aligning, every pair of bases has a score value (may be negative), the ultimate score is the sum of them. Matching score value is greater than mismatching score value. In addition, the mismatching pair, T and G, has greater score value than other mismatching pairs. It is called a wobble. Insertions or deletions may occur when aligning, and they are called, represented by "-". Every gap has a penalty (negative sore value), and an open gap has smaller penalty then an extensive one. Similar to targetScan, miRanda also defines seeds. Gaps at seeds is not allowed, and positive score values and penalty score values at seeds are several times more than ones at non-seed areas.

References

[1] Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. Benjamin P Lewis, Christopher B Burge, David P Bartel. Cell, 120:15-20 (2005).

[2] Most Mammalian mRNAs Are Conserved Targets of MicroRNAs. Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, David P Bartel. Genome Research, 19:92-105 (2009).

[3] MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing Andrew Grimson, Kyle Kai-How Farh, Wendy K Johnston, Philip Garrett-Engele, Lee P Lim, David P Bartel. Molecular Cell, 27:91-105 (2007).

[4] Weak Seed-Pairing Stability and High Target-Site Abundance Decrease the Proficiency of lsy-6 and Other miRNAs. David M Garcia, Daehyun Baek, Chanseok Shin, George W Bell, Andrew Grimson, David P Bartel Nat Struct Mol Biol, 18:1139-1146 (2011).

[5] Approximate matching of regular expressions. Myers EW and Miller W, Bull Math Biol, 1989; 51(1):5-37.

Appendix: conservation across species and computing PCT

The genetic relationship between species are represented by a phylogenetic tree, a node of which is a species. Every node has a number, indicating the branch length at this of the node. TargetScan uses 11 phylogenetic trees, one of which is obtained from considering all genes. The other 10, which are numbered from 1 to 10, are obtained from a gene group, after dividing all genes into several groups. Every tree has a threshold, and the tree numbered greater has a greater threshold.

In order to obtain conservation information, a reference species must first be selected. For any gene, if a position at the reference sequence happens to have a base, not "-", and at least one of the other sequences has a base at the same location, then this position's number(offset) is mapped to a list consisting of all the species(including reference species), from which the gene sequences are obtained. For every mapped species list, from the overall phylogenetic tree an overall branch length can be computed. The median of all overall branch lengthes is this gene's

branch length. When this branch length is not less than the threshold of the nth partial phylogenetic tree but less than the threshold of the (n+1)th tree, the nth tree is the one used to determine the conservation across species.

Select the numbered partial phylogenetic tree, which is determined by the field Gene_id. If the value at Site_type field is the same as the one at Group_type field, then all seed combination types of this group is the same, so another branch length can be computed from this tree and the species list of the field Species_in_this_group. Every seed type also has a threshold. If the preceding branch length is no less than this threshold, then the gene is considered to be conservative across species. But if the values at fields Sity_type and Group_type are different, then the species list at field Species_in_this_group_with_this_site_type is selected. If the seed combination type is 7mer-m8 or 7mer-1a, then the branch length is computed and use it to determine whether the gene is conservative. Otherwise if the seed combination type is 8mer-1a, then the branch length is also computed. If this length is no less than the 8mer-1a threshold, then the gene is conservative, too. If the reverse is true and the seed combination type is also 7mer-1a or 7mer-m8, then the same process is performed as if the combination type were 7mer-1a or 7mer-m8. The reason is that if a seed combination type is 8mer-1a, then it is also 7mer-m8 or 7mer-1a, and the thresholds of the tree seed type is increasing, so the three steps must be taken in order.

The PCT is computed by the formula:

$$\mathrm{PCT} = \max(0, \beta_0 + \frac{\beta_1}{1 + \exp(-\beta_2 x + \beta_3)})$$

In which x is the seed type branch length, β_0、β_1、β_2 and β_3 are coefficients corresponding to some seed type. Their values are obtained by the method of least square estimation considering huge amounts of samples. If the seed type is 8mer-1a, then it can be seen as one or two other types, so multiple PCT values can be obtained. The ultimate one is the biggest of them.