

## Master de Bioinformatique et Biostatistiques 2ème année

Examen de Combinatoire et Algorithmique des Structures Moléculaires  
partie « Algorithmique »

mardi 28 novembre 2006

Durée : 1h30. Tous documents sont autorisés.

Le travail proposé se fonde sur l'article suivant, paru dans *Genes and Development*, 2003.

## The microRNAs of *Caenorhabditis elegans*

Lee P. Lim,<sup>1,2,3,4</sup> Nelson C. Lau,<sup>1,2,3</sup> Earl G. Weinstein,<sup>1,2,3</sup> Aliaa Abdelhakim,<sup>1,2,3</sup> Soraya Yekta,<sup>1,2</sup>  
Matthew W. Rhoades,<sup>1,2</sup> Christopher B. Burge,<sup>1,5</sup> and David P. Bartel<sup>1,2,6</sup><sup>1</sup>Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA, and <sup>2</sup>Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA

Dans cet article, les auteurs développent un programme, miRscan, pour la détection des gènes de miRNA dans le génome du nématode *C. elegans*. Les gènes sont repérés notamment par la présence du précurseur constitué d'une longue tige-boucle (hairpin) d'environ 70nt. Nous nous intéresserons à la partie bioinformatique de l'étude, présentée ainsi:

### *Computational prediction of C. elegans miRNA genes*

We developed a computational tool to specifically identify miRNAs that are conserved in two genomes and have the features characteristic of known miRNAs. To identify miRNAs in nematodes, the *C. elegans* genome was first scanned for hairpin structures with sequences that were conserved in *Caenorhabditis briggsae*. About 36,000 hairpins were found that satisfied minimum requirements for hairpin structure and sequence conservation. This procedure cast a sufficiently wide net to capture 50 of the 53 miRNAs previously reported to be conserved in the two species [Lau et al. 2001; Lee and Ambros 2001]. These 50 published miRNA genes served as a training set for the development of a program called MiRscan, which was then used to assign scores to each of the 36,000 hairpins, evaluating them based on their similarity to the training set with respect to the following features: base pairing of the miRNA portion of the fold-back, base pairing of the rest of the fold-back, stringent sequence conservation in the 5' half of the miRNA, slightly less stringent sequence conservation in the 3' half of the miRNA, sequence biases in the first five bases of the miRNA (especially a U at the first position), a tendency toward having symmetric rather than asym-

metric internal loops and bulges in the miRNA region, and the presence of two to nine consensus base pairs between the miRNA and the terminal loop region, with a preference for 4–6 bp (Fig. 1A).

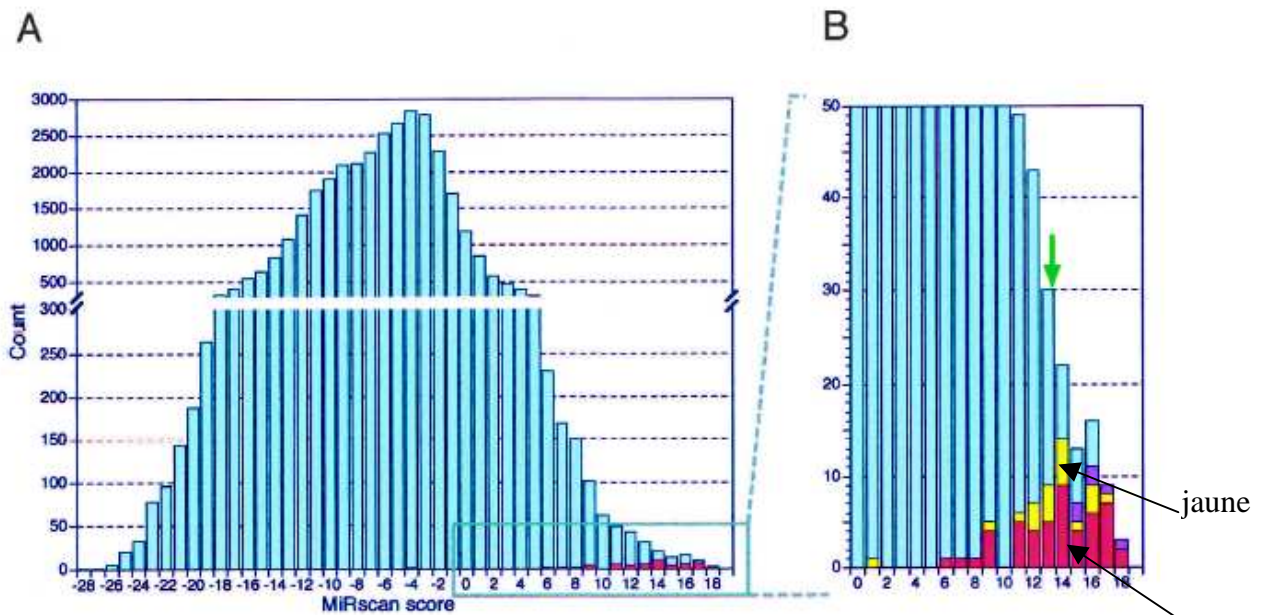
The distribution of MiRscan scores for the ~36,000 hairpins illustrated the ability of MiRscan to discern the 50 miRNA genes of the training set, which fell mostly in the high-scoring tail of the distribution (Fig. 2). Of the features evaluated by MiRscan, base-pairing potential and sequence conservation played primary roles in distinguishing known miRNAs (Fig. 1B). Some of the other conserved hairpins also scored highly; 35 had scores exceeding 13.9, the median score of the 58 known miRNAs (Fig. 2B). These 35 hairpins were carried forward as the top miRNA candidates predicted by MiRscan.

Questions :

1.1 Quels sont les principaux critères utilisés pour la prédiction ?

1.2 A quoi sert le génome de *C. Briggsae* ? Que se passerait-il si on n'avait pas accès à ce génome ?

Les scores miRscan des 36000 hairpins trouvées se distribuent ainsi :



**Figure 2.** Computational identification of miRNA genes. (A) The distribution of MiRscan scores for 35,697 *C. elegans* sequences that potentially form stem loops and have loose conservation in *C. briggsae*. Note that the Y-axis is discontinuous so that the scores of the 50 previously reported miRNA genes that served as the training set for MiRscan can be more readily seen [red]. Scores for these 50 genes were jackknifed to prevent inflation of their values because of their presence in the training set. (B) An expanded view of the high-scoring tail of the distribution. This view captures 49 of the 50 genes of the training set (red). The median score of the 58 previously reported miRNA loci that satisfy the current criteria for designation as miRNA genes [Ambros et al. 2003] is 13.9 (green arrow). Note that this median score was the midpoint between the scores of the 29th and 30th highest-scoring loci of the 50-member training set; namely, it was designated the median score after including the 8 previously reported miRNA genes that were not in the training set because they were lost during the identification of conserved hairpins, usually because they lacked sufficient *C. briggsae* homology. Scores of genes validated by cloning are indicated (yellow), as are scores of six genes that have not yet been cloned but were verified by Northern analysis (purple). (C) Examples of miRNA genes identified by MiRscan with the Northern blots that served to

Lisez la légende correspondant aux parties A et B et de la figure. Le « Jackknife » est l'opération qui consiste à retirer un élément du training set avant de calculer son score, de façon à ce que ce score ne soit pas artificiellement gonflé par sa présence dans le training set.

2.1 Qu'est-ce que le « training set » ?

2.2 Quels sont les aspects positifs et négatifs de la distribution des scores des vrais miRNA ?

Le fonctionnement de miRscan est détaillé ci-dessous. Les 40,000 hairpins mentionnées au début sont celles produites par la combinaison prédiction de structure + conservation. Le « 21-nt miRNA » est la partie du précurseur qui sera excisée pour former le miRNA mature.

#### MiRscan

Of the -40,000 pairs of hairpins, 35,697 had the minimal conservation and base pairing needed to receive a MiRscan score. Among this set were 50 of the 53 previously published miRNAs that were reported to be conserved between *C. elegans* and *C. briggsae* (Lau et al. 2001; Lee and Ambros 2001). [miR-53 is included as a previously reported conserved miRNA because it is nearly identical to miR-52, which has a highly conserved *C. briggsae* ortholog (Lau et al. 2001; Lee and Ambros 2001). The three conserved genes missing from the -36,000 pairs of hairpins were *mir-56*, *mir-75*, and *mir-88*. The reverse complements of *mir-75* and *mir-88* were later observed among the -36,000 hairpins and given scores (Table 1).] The MiRscan program was developed to discriminate these 50 known miRNA hairpins from background sequences in the set of -36,000 hairpins. For a given 21-nt miRNA candidate, MiRscan makes use of the seven features derived from the consensus hairpin structure illus-

trated in Figure 1A:  $x_1$ , "miRNA base pairing," the sum of the base-pairing probabilities for pairs involving the 21-nt candidate miRNA;  $x_2$ , "extension of base pairing," the sum of the base-pairing probabilities of the pairs predicted to lie outside the 21-nt candidate miRNA but within the same helix;  $x_3$ , "5' conservation," the number of bases conserved between *C. elegans* and *C. briggsae* within the first 10 bases of the miRNA candidate;  $x_4$ , "3' conservation," the number of conserved bases within the last 11 bases of the miRNA candidate;  $x_5$ , "bulge symmetry," the number of bulged or mismatched bases in the candidate miRNA minus the number of bulged or mismatched bases in the corresponding segment on the other arm of the stem loop;  $x_6$ , "distance from loop," the number of base pairs between the loop of the stem loop and the closest end of the candidate; and  $x_7$ , "initial pentamer," the specific bases at the first five positions at the candidate 5' terminus.

For a given feature  $i$  with a value  $x_i$ , MiRscan assigns a log-odds score

$$s_i(x_i) = \log_2 \left( \frac{f_i(x_i)}{g_i(x_i)} \right),$$

where  $f_i(x_i)$  is an estimate of the frequency of feature value  $x_i$  in miRNAs derived from the training set of 50 known miRNAs, and  $g_i(x_i)$  is an estimate of the frequency of feature value  $x_i$  among the background set of -36,000 hairpin pairs. The overall score assigned to a candidate miRNA is simply the sum of the log-odds scores for the seven features:

$$S = \sum_{i=1..7} s_i(x_i).$$

#### Questions :

- 3.1 Proposer une explication au fait que 3 gènes de miRNA conservés (*mir-56*, *mir-75* et *mir-88*) soient absents des 36000 candidats analysés.
- 3.2 Proposez au moins deux raisons pour lesquelles certaines régions du miRNA peuvent être particulièrement conservées.
- 3.3 Justifiez le choix de la formule de calcul de score ( $\log(f_i(x_i)/g_i(x_i))$ )

