On the length of mature microRNAs

Dave Tang
RIKEN Yokohama

Derek de Rie
VU University Amsterdam

June 10, 2014

Abstract

Despite their size, microRNAs (miRNAs) can have dramatic effects on human health due to their direct effect on transcripts. The misregulation of miRNAs can cause a wide spectrum of diseases, such as cancer, and as such they are under heavily investigation.

1 Introduction

MicroRNAs (miRNAs) were discovered in 1993[1] and are the most well known class of non-coding RNAs (ncRNAs).

2 Methods and results

All code underlying this work is available at `https://github.com/davetang/mirna_length`. Briefly, random sequences were generated using R (version 3.1.0) and the R Bioconductor package Biostrings[2]. We generated three sets of one million random sequences that ranged from 15 to 30 base pairs in length, totalling 48 million sequences. The first two sets of sequences were generated based on the multinomial sequence model where each nucleotide in the sequence is independent and identically distributed. The first set of randomly generated sequences used an equal probability for each nucleotide ($p_a = 0.25$, $p_c = 0.25$, $p_g = 0.25$, $p_t = 0.25$). The second set used probabilities based on the what was observed in the human genome (hg38): $p_a = 0.29$, $p_c = 0.20$, $p_g = 0.21$, $p_t = 0.30$. The last set of randomly generated sequences used a Markov chain model, where the next sequence depends on the previous sequence. Transitions probabilities (see figure 1) were derived on dinucleotide frequencies of mature human miRNAs that were downloaded from miRBase[3]. The probability of the first base was derived from the frequency of nucleotides at the first base of the human mature miRNAs. The alignment of the sequences was performed using BWA[4] (version 0.7.9a-r786) using aln/samse and summaries were created using Perl scripts.

Bar plot of mapped versus unmapped at each sequence length for the three sets.

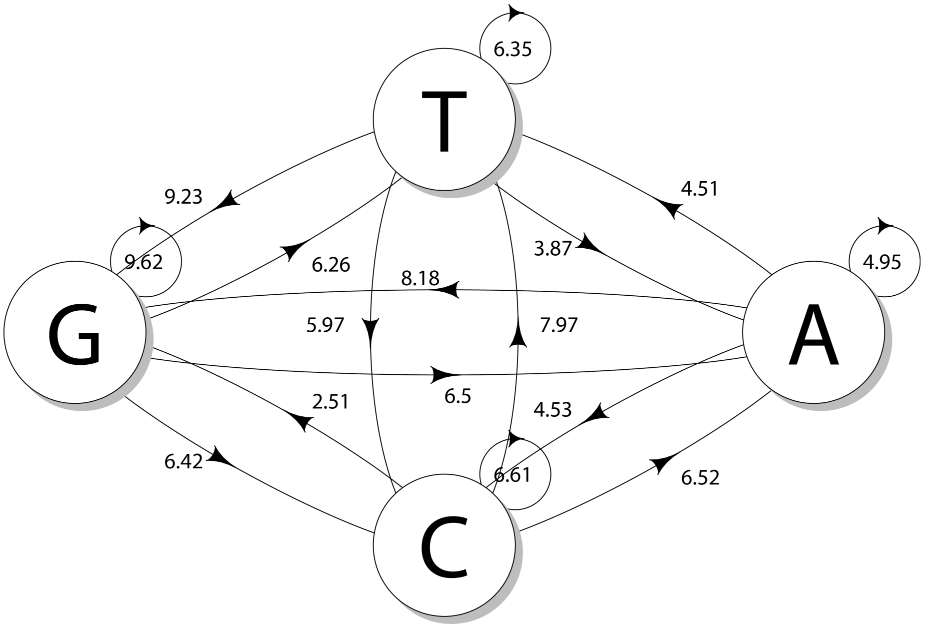

Figure 1: Transition diagram based on dinucleotide frequencies of miRBase human miRNAs.

Bar plot of perfectly mapped at each sequence length for the three sets.

3 Discussion

4 Conclusions

Mammalian miRNAs are able to recognise their target miRNA by as little as 6-8 nucleotides; this region is known as the seed region, which lies at the 5' end of a miRNA. However despite this, most miRNAs are 22-23 base pairs in length.

References

[1] R. C. Lee, R. L. Feinbaum, and V. Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, Dec 1993.

[2] H. Pages, P. Aboyoun, R. Gentleman, and S. DebRoy. *Biostrings: String objects representing biological sequences, and matching algorithms.* R package version 2.30.1.

[3] A. Kozomara and S. Griffiths-Jones. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, 39(Database issue):D152–157, Jan 2011.

[4] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009.