Computational Biology and Chemistry 32 (2008) 222–226

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Brief Communication

On the relationship between GC content and the number of predicted microRNA binding sites by MicroInspector

Nicole Davis[a], Natasha Biddlecom[a], David Hecht[a], Gary B. Fogel[b,*]

[a] Southwestern College, 900 Otay Lakes Road, Chula Vista, CA 91910, USA
[b] Natural Selection Inc., 9330 Scranton Road, Suite 150, San Diego, CA 92121, USA

ARTICLE INFO

Article history:
Received 11 February 2008
Received in revised form 15 February 2008
Accepted 16 February 2008

Keywords:
MicroInspector
MicroRNA
Target prediction
GC content

ABSTRACT

MicroRNA GC content and length is believed to play a role in the prediction of putative microRNA targets. MicroInspector was evaluated to determine the extent to which these characteristics of microRNAs play a role in binding site predictive accuracy. A strong bias towards under predicting the number of expected bindings sites for low GC content sequences was observed, especially for microRNAs with <50% GC content. Researchers working with organisms with unusually low GC content should be aware of this bias.

1. Introduction

MicroRNAs (miRs) are known to play key roles in cellular differentiation and regulation, and operate via basepairing interactions with a target sequence (Foshay and Gallicano, 2007; Marquez and McCaffrey, 2007; Liu et al., 2008). miRs are also believed to be involved with tumorigenesis and some miRs may have diagnostic value (Lowery et al., 2008; Schetter et al., 2008). Unfortunately the number of experimentally verified miRs exceeds the number of experimentally verified miR targets (Betel et al., 2008). Computational and experimental approaches to help to ascertain putative miR targets are highly desired by the research community.

Several bioinformatics approaches have been generated for the prediction of miR targets in sequence information, based on databases of known miRs and calculations of sequence similarity and/or free energy of binding (Sethupathy et al., 2006a,b; Thadani and Tammi, 2006). For example, MicroInspector (http://mirna.imbb.forth.gr/microinspector/) is a tool for the detection of microRNA binding sites (Rusinov et al., 2005). While there are many tools that attempt to identify as many mRNA targets as possible for each given miR, MicroInspector attempts to identify possible binding sites based on rules for the free energy of binding that are appropriate to known miRs for each organism of interest. These putative binding sites can be examined with follow-on experimentation (Bonnet et al., 2004a,b).

The authors of MicroInspector acknowledged several possible issues with the reliance on free energy calculation for binding prediction. These included a higher likelihood for binding site prediction with increase in GC content of the sequence and miR length (Rusinov et al., 2005). We chose to better understand the relationship of GC content and miR length to binding site prediction from MicroInspector, specifically to determine which organisms might be most affected by such a bias is simply because of their genomic GC content.

2. Materials and Methods

2.1. MicroInspector

MicroInspector (Rusinov et al., 2005) is a web-based tool for searching miR binding sites in a target RNA sequence. The user is first asked to input a sequence to be analyzed (which can be either RNA or DNA). The user can either input the sequence data manually or provide GenBank accession numbers. A hybridization temperature can then be adjusted by the user if required. This value affects the calculation of binding affinity and the program authors recommend the use of a default temperature of 37 °C. In addition, a value for a minimum free energy also needs to be defined, and the default value is −20 kcal/mol. This cut-off affects the number of results that are presented to the user: only those results that have lower energy than the cut-off are displayed. As a final step, the user must select an miR database to be interrogated, effectively allowing the user to choose a miR database in an organism-specific manner.

* Corresponding author. Tel.: +1 858 455 6449; fax: +1 858 455 1560.
E-mail address: gfogel@natural-selection.com (G.B. Fogel).

1476-9271/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2008.02.004

Once these parameters are established, the algorithm scans each target sequence for every miR sequence from the chosen miR database consecutively in an attempt to identify possible regions of hybridization. This is accomplished using a window of 6 nt sliding by 1 nt throughout the sequences. For each window pair, the approach searches for domains having five Watson–Crick base pairs or four Watson–Crick pairs with one additional G:U at any location in the window. If these conditions are not found, the window is shifted by 1 nt and the calculation is repeated until all windows have been examined. If any window does satisfy the basepairing criteria, then additional methods are used to examine hybridization and folding. The output of this approach to the user is a list of putative binding locations of miRs to any sequence that was input to the original query window.

For the purposes of our investigation, all human and random sequences as well as their reverse complements were processed through MicroInspector (Rusinov et al., 2005) to identify the number and location of possible microRNA binding sites. Standard default settings were used as recommended.

2.2. Sequence Curation

One hundred random sequences each of length 100 nucleotides were generated using the Sequence Manipulation Suite (Stothard, 2000) with default settings. This process was repeated nine times with each set using a different GC content in the range (20–80%) to cover reasonable estimates of GC content for genomic sequence information as well as known microRNAs. Hairpin sequences for all known human, *Caenorhabditis briggsae*, *Arabidopsis thaliana*, and *Rattus norvegicus* microRNAs were downloaded from the microRNA registry (Griffiths-Jones, 2004; Griffiths-Jones et al., 2006) and for each sequence, the GC content was calculated. MicroRNAs were binned into three GC content categories (low ≤33%GC, medium 34–66%GC, and high ≥67%GC).

3. Results

3.1. Human MicroRNAs and Random Sequences

Analysis of all known human microRNA hairpin sequences in the miR registry ($n = 533$) indicates that GC content is normally distributed with a mean of 50.52% and standard deviation of 10.35% (Fig. 1). The mean hairpin length is 89 nucleotides with a standard deviation of 13 nucleotides and is not normally distributed (Fig. 2). The microRNA GC content is similar to the estimated GC content of the human genome; the mean GC content of the human genome has been estimated to be ~40% (Lander et al., 2001). When human microRNA hairpin sequences are provided to MicroInspector, the number of binding sites predicted by MicroInspector was found to be dependent on the GC content of the sequence being provided.

To further analyze this behavior, the set of known human microRNAs was subdivided into categories of GC content by arbitrarily chosen three equal bins based on GC% (Table 1). The microRNAs used for these different GC contents are provided in

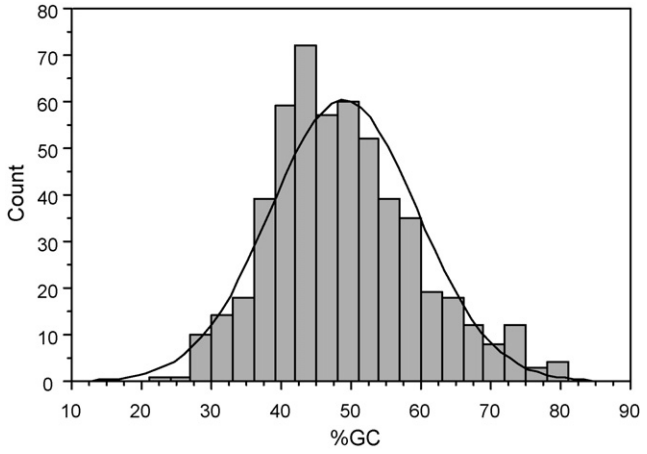

Fig. 1. Distribution of GC content (%) for known human microRNA hairpin sequences in the microRNA registry ($n = 533$). According to the Kolmogorov–Smirnov test, the data are not likely to be normally distributed ($P = 0.00$ where the normal distribution has mean = 50.52 and standard deviation = 10.35).

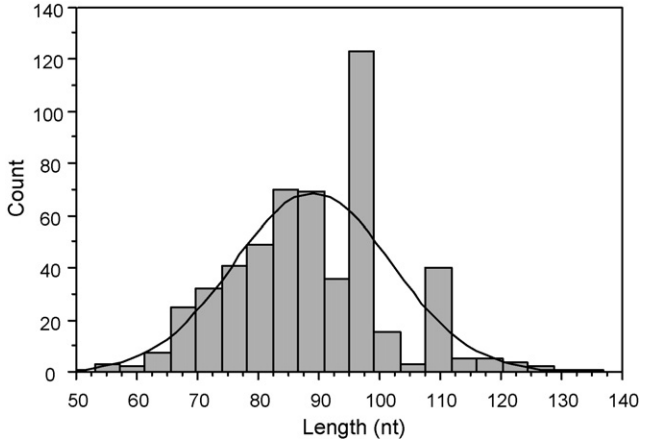

Fig. 2. Distribution of length in nucleotides (nt) for known human microRNA hairpin sequences in the microRNA registry ($n = 533$). According to the Kolmogorov–Smirnov test, the data are not normally distributed ($P = 0.00$ where the normal distribution has mean = 90.85 and standard deviation = 13.61). In particular there is a large overabundance of sequences with length 95–100 nt.

Table 2. Sequences with a high mean GC content yielded roughly 10 times the mean number of predicted binding sites using MicroInspector when compared to those with a low mean GC content, however, the mean number and standard deviation of predicted binding sites for sequences in the medium GC content and high GC content categories were quite similar. The observation of such a dramatic difference in the number of predicted binding sites by MicroInspector relative to GC content was unanticipated, especially due to the observed variance in GC content for known human microRNA hairpins in the miR registry. Using an unpaired *t*-test, this difference in the number of predicted binding sites was statis-

Table 1
Categorization of microRNA hairpin sequences by GC content (low, medium, or high) with resulting mean GC content by category and number of microRNA binding sites predicted by MicroInspector

GC category	Number of sequences	Mean GC%	Predicted number of binding sites by MicroInspector ($\mu \pm \sigma$)
Low	26	30.1	1.3 ± 1.5
Medium	30	50.0	8.8 ± 5.4
High	30	71.5	10.0 ± 4.9

The predicted number of binding sites by MicroInspector in the GC category "low" is statistically different from either the "medium" ($P < 0.0001$) or "high" GC categories $P < 0.0001$).