Implementation of K-Factor Algorithm and Its Application to the Identification of miRNA Regulatory Elements

Ji Lee, Department of Bioengineering, Pennsylvania State University

Bino John, Mentor, Department of Computational Biology, University of Pittsburgh School of Medicine

Specific Aims

MicroRNAs (miRNAs) are non-protein-coding sequences of approximately 22 nucleotide long RNA that are believed to negatively regulate gene expression[1,2]. Although approximately 400 human miRNAs have been discovered, their function remains largely unknown. MiRNAs are thought to prevent mRNA translation by base pairing to the 3' UTR of the target mRNAs, causing the silencing of the target gene[2]. The precise mechanisms of miRNA-mediated gene regulation and the cascade of molecular events that lead to the biogenesis of miRNAs are beginning to emerge.

Evidence suggests that there is a complex network of transcriptional and post-transcriptional events that regulate the expression of miRNA genes. A study done using chromatin immunoprecipitation and DNA microarrays suggests that at least one of the known transcription factors OCT4, SOX2, and NANOG regulate the transcription of multiple miRNAs[3]. Sequence elements upstream of the miRNAS *miR-1, miR-223,* and *miR-17* are also known to be affected by certain transcription factors[4,5,6]. A computational scan of the upstream regions of nematode miRNAs suggests that a specific sequence motif is present in the upstream regions of almost all transcribed nematode miRNAs[7]. Hence, it is reasonable to postulate that a number of miRNA transcription factors are sequence-specific and bind to miRNA promoters at select sites. Building

upon this hypothesis, a computational method, K-Factor, was developed to accurately identify regulatory elements that regulate the expression of miRNAs (John et al, manuscript in preparation).

We propose to implement the K-Factor algorithm as a module within a flexible application framework written in Java. The completed K-Factor implementation will be a highly extensible, platform-independent desktop application. The software will be used to identify regulatory sequence motifs that regulate the expression of miRNAs. The accuracy of the proposed implementation will be evaluated by comparing the predicted sequence motifs to the results of the aforementioned in-house study.

Method

K-Factor Algorithm

The K-Factor computational method is designed to identify regulatory sequences of length k ("k-mers") that are embedded in a given set, S of user-defined DNA sequences of a specific genome, G. For a given number of user-defined sequences ($n_{S,G}$), K-Factor involves the following steps: (1) *m* sets of sequences (reference sets, R) are extracted from random locations in G. Each set in R consists of $n_{S,G}$ sequences that have identical lengths to the sequences in S; (2) the sequence densities of all possible k-mers in S and each of the *m* sets in R are calculated. The density is defined as the ratio of the number of occurrences of a given k-mer to the total number of nucleotides in S; (3) A score that reflects the bias of each k-mer to preferentially occur in S with respect to each reference set, R_i is determined. The enrichment score for a given k-mer with respect to R_i is defined as its density ratio in S to R_i; (4) The K-Factor score for each k-mer is

computed as the average enrichment score over all sequence sets in R; and (5) A list of k-mers with K-factor scores above a user-defined threshold of *T* is extracted.

Implementation of K-Factor

The K-Factor method will be implemented as a plug-in module within the construct of a modular, extensible application framework. The application and its framework will be written in Java 2 SE 5, while possibly being backwards compatible to Java 2 SE 1.4. It will be platform-independent and will initially be built as a skeletal structure with a rudimentary implementation of the K-Factor algorithm.

The roadmap for the first release calls for an optimized K-Factor to be implemented as part of a robust motif identification package. The package will offer access to all parameters of the K-Factor algorithm, top to bottom, and the abilities to tweak them accordingly. A full suite of input and output capabilities will also be implemented, providing the facilities to read both DNA and protein sequences, parse various sequence formats like FASTA, extract sequences from various genomic databases, and output to various file types, among other potential functionality.

A visualization and graphical analysis package is scheduled for future releases, offering the capabilities to perform various tasks such as frequency analysis and mapping regulatory networks (with the aid of the application Cytoscape[8]). This module will essentially be linked to the motif identification module, providing easy channeling for K-Factor output into a wide array of graphical analysis options.

The application framework will support the integration of other algorithms and visualization options within its modular architecture. Extensibility will not be limited to

just model functionality, as user-interface (UI) implementations will be pluggable themselves, thus allowing interchangeability of different UI options.

Summary

K-Factor is a useful method for the identification of genomic sequence motifs that are potential regulatory elements. A flexible, extensible implementation of K-Factor within a larger application framework will allow K-Factor to be tested, tweaked, and demonstrated. The proficiency of K-Factor will be demonstrated by its application to known human miRNA sequences. The extensibility of the proposed system will ensure that future tools and functionality can be easily added to improve the K-Factor method and the underlying framework itself.

References

1. Pasquinelli, A. E., Hunter, S. & Bracht, J. MicroRNAs: a developing story. *Curr. Opin. Genet. Dev.* **15**, 200-205 (2005).

2. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281-297 (2004).

3. Boyer, L.A., Lee T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., Gifford, D.K., Melton, D.A., Jaenisch, R., and Young, R.A. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 828-30 (2005).

4. O'Donnell, K. A., Wentzel, E. A., Zeller, K. I., Dang, C. V. & Mendell, J. T. c-Myc-regulated microRNAs modulate E2F1 expression. *Nature* **435**, 839-843 (2005).

5. Zhao, Y., Samal, E. & Srivastava, D. Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature* **436**, 214-220 (2005).

6. Fazi, F., Rosa, A., Fatica, A., Gelmetti, V., De Marchis, M. L., Nervi, C. & Bozzoni, I. A minicircuitry comprised of microRNA-223 and transcription factors NFI-A and C/EBPalpha regulates human granulopoiesis. *Cell* **123**, 819-831 (2005).

7. Ohler, U., Yekta, S., Lim, L. P., Bartel, D. P. & Burge, C. B. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA.* **10**, 1309-1322 (2004).

8. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks." *Genome Res.* **13.11** (2003): 2498-504.