



S2-1

Genome-wide Prediction of Intronic MicroRNAs

Seungyoon Nam, Bumjin Kim and Sanghyuk Lee

Division of Molecular Life Sciences, Ewha Womans University, Seoul 120-750, Korea

MicroRNAs have been known to regulate gene expression at the post-transcriptional level across a broad range of organisms from plants to human. We describe a computational method that predicts novel intronic microRNAs using machine-learning approaches. Careful examination of known microRNAs reveals that approximately 70% of them originate from introns of known (protein-coding) genes. A computational pipeline was developed to identify novel intronic microRNAs using phylogenetic conservation and structural features of pre-miRNAs. First, we made a genome-wide catalog of conserved stem-loop sequences in introns of known genes. The secondary structure of intronic sequence (80 bp long) was predicted using the Vienna RNA package. Then, we extracted ~30 features that would be relevant to microRNA prediction, which could be classified into five subjects (i) phylogenetic shadowing, (ii) structural features of pre-miRNA, (iii) stability of secondary structures tested by random permutation, (iv) sequence composition modeled using the first Markov chain, and (v) genomic neighborhood. Thirty features were tested using known intronic microRNAs in the Rfam database, and a decision tree was constructed with significant features. Our analysis gave ~3,000 candidate microRNAs. The result is compared with the Rfam and result by Plasterk and coworkers (Berezikov et al. Cell 2005).