Automatic extraction of hairpin sequences from genome-wide data

C.A. Yones[1], G. Stegmayer[1], L. Kamenetzky[2], D.H. Milone[1]

1 Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL, CONICET, Ciudad Universitaria UNL, (3000) Santa Fe, Argentina.
2 Instituto de Investigaciones en Microbiología y Parasitología Médica (UBA), CONICET, Paraguay 2155, piso 13 (1121), Buenos Aires, Argentina.

Background

Extracting stem-loop sequences from raw genome-wide data is very important for some data mining tasks in bioinformatics. For example, to discover new microRNA precursors (pre-miRNA) in genome-wide data using machine learning techniques, which need as input the stem-loop sequences to build a prediction model. The genome preprocessing is very important because it has a strong influence on the prediction results. All well-known pre-miRNAs must be found in the resulting sequences, thus all hairpins must be properly trimmed. Although there are some scripts that can be adapted and put together to achieve this task, to the best of our knowledge they are outdated and do not take advantage of the latest advances in secondary structure prediction. For example, the scripts from mirCheck for searching hairpins use Einverted, developed in 1999.

Results and conclusions

We have developed a simple and integrated tool[1] that automatically extracts and folds all hairpin sequences from raw genome-wide data. It predicts the secondary structure of several overlapped segments of the raw genome, with longer length than the mean of well-known pre-miRNAs of the species under processing, ensuring that no pre-miRNA is inappropriately cut. Stem-loops that meet specified requirements are extracted and trimmed. Each genome can be processed in parallel and the implementation is memory efficient since it can automatically split large multifasta files. Several genomes were processed and the results were compared to those from mirCheck (Table 1). The number of hairpins and known pre-miRNAs found were significatively higher for the proposed method

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
C. Yones, G. Stegmayer, L. Kamenetzky & D. H. Milone; "Automatic extraction of hairpin sequences from genome-wide data"
4th ISCB-LA Bioinformatics Conference, nov, 2016.

[1] Freely available at: https://sourceforge.net/projects/sourcesinc/files/hextractor/.

Table 1

Species	Hairpins extracted		pre-miRNAs founded		Known pre-miRNAs
	mirCheck	Proposed tool	mirCheck	Proposed tool	
Anopheles gambiae	1,410,532	4,276,254	92.42 %	100 %	66
Caenorhabditis elegans	875,588	1,739,124	90.00 %	99.60 %	250
Danio rerio	11,028,128	23,214,338	93.06 %	99.42 %	346
Echinococcus multilocularis	509,530	1,898,911	81.81 %	100 %	22

Supported by: CONICET, MinCyT, UNL.

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
C. Yones, G. Stegmayer, L. Kamenetzky & D. H. Milone; "Automatic extraction of hairpin sequences from genome-wide data"
4th ISCB-LA Bioinformatics Conference, nov, 2016.