Supplementary Material

Deep Learning for the discovery of new pre-miRNAs: helping the fight against COVID-19

L. A. Bugnon, J. Raad, G. Merino, C. Yones, F. Ariel, D.H. Milone, G. Stegmayer

Table S1: Features computed for each hairpin sequence.

Feature name	Description
nt_proportion	Ratio of each base in the sequence (A, C, G and T).
dinucleotide_proportion	Ratio of dinucleotide elements of each kind, making 16 features for the possible binary combinations of the 4 nucleotides.
gc_content	Proportion of guanine-cytosine on the sequence.
gc_ratio	Ratio between guanine and cytosine.
sequence_length	Length of the sequence.
stem_number	Number of stem-loops.
avg_bp_stem	Average of nucleotides per stem.
longest_stem_length	Longest region where the pairing is perfect.
terminal_loop_length	Number of nucleotides in the stem region.
bp_number	Number of base-pairs.
dP	Number of base pair divided by the nucleotide number.
bp_proportion	Number of each possible base pair normalized by sequence length.
bp_proportion_stem	Proportion of base pairs on stems.
triplets	Frequencies of secondary structure triplets, this is the 32 possible combinations of the 4 nucleotides in a sequence of 3.
MFE	Minimum free energy.
EFE	Normalized Ensemble Free Energy calculated with RNAfold (-p option).
ensemble_frequency	Frequency of the minimum free energy in the ensemble.
diversity	Structural diversity calculated with RNAfold (-p option).
mfe_efe_difference	Calculated as $\|MFE - EFE\|/l$.
dQ	Calculated as $1/L \sum_{i<j} p_{ij} log_2 p_{ij}$, where L is length and p_{ij} is the probability of pairing of nucleotides i and j.
dG	Minimum free energy divided by sequence length.
MFEI1	Ratio between the minimum free energy and the $\%C + G$.
MFEI2	dG/N_s, where N_s is the number of stems.
MFEI4	MFE/N_b, where N_b is the total number of base pairs in the secondary structure.