SciForum
MOL2NET

Prediction of mRNA expression in cow's milk using mRNA secondary structures and Machine Learning classifiers

Rodrigo Martín [1,*], Yong Liu [1,2], Omar Landaeta [1], Luis Felipe Llamas [1], Chuanshe Zhou [2,3], Zhiliang Tan [2,3], Haibo Zhang [4], Cristian R Munteanu [1,*]

[1] Computer Science Faculty, University of A Coruna, Campus de Elviña s/n, 15071 A Coruña, Spain; E-Mails: r.martin1@udc.es (R.M.); y.liu86@outlook.com (Y.L.); omarlandaeta@gmail.com (O.L.); lfllamas93@gmail.com (L.F.L.); c.munteanu@udc.es (C.R.M.)

[2] Key Laboratory for Agro-Ecological Processes in Subtropical Region, Hunan Research Center of Livestock and Poultry Sciences, South Central Experimental Station of Animal Nutrition and Feed Science in the Ministry of Agriculture, Institute of Subtropical Agriculture, The Chinese Academy of Sciences, Changsha, Hunan 410125, P.R. China; E-Mails: y.liu86@outlook.com (Y.L.); zcs@isa.ac.cn (C.Z.); zltan@isa.ac.cn (Z.T.)

[3] Hunan Co-Innovation Center of Animal Production Safety, CICAPS, Changsha, Hunan 410128, P.R. China; E-Mails: zcs@isa.ac.cn (C.Z.); zltan@isa.ac.cn (Z.T.)

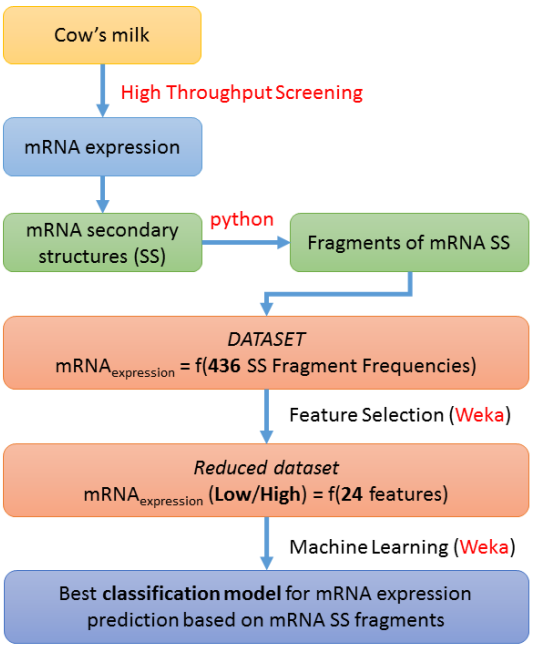
[4] College of Life Science and Environmental Resource, Yichun University, Jiangxi Yichun, 336000, China; E-Mail: zhanghaiboainide@163.com (H.Z.)

***** Author to whom correspondence should be addressed; E-Mail: c.munteanu@udc.es; Tel.: +34 981167000x1302; Fax: +34 981167160

Abstract: The mRNA molecules expressed in cow's milk are important molecular biomarkers for different physiological and pathological conditions in cattle. The prediction of the quantity that a specific mRNA type could be expressed in cow's milk is a challenging theoretical task. The current study presents for the first time several different Machine Learning models to predict the mRNA expression using the mRNA secondary structure fragments. This unique methodology is based on a dataset of experimental mRNA expression data. Each mRNA molecule has a specific secondary structure represented as a string that can be used to read all the possible mRNA secondary structure fragments. This information is used as input for the Machine Learning methods from Weka software in order to obtain classification models that can predict low and high expression of new mRNA types in the cow's milk. The mRNA expression levels have been measured with High Throughput Screening techniques. The initial features included the counting of the mRNA secondary structure fragments for each expressed mRNA. The model features were transformed in frequencies and the expression levels were converted into low and high classes. In order to reduce the high number of possible features, a feature selection method has been applied. Thus, the best classification model was obtained with BayesNet method and is based on 24 features and 4067 cases. The model has the true positive rate for the low mRNA expression class of 0.78 (average true positive rate of 0.66). Further studies are needed improve the current results, using datasets with different feature sets and more advanced Machine Learning methods.

Keywords: *mRNA secondary structures, Machine Learning classifiers, mRNA expression*

Graphical Abstract:



Introduction: The mRNA expression in cow's milk is an important biomarker for the cattle conditions [1,2]. The current study proposes a method to predict the low or high expression levels of mRNA using mRNA secondary structure fragments and Machine Learning classifiers [3].

Materials and Methods: In the first step, the mRNA expression levels were measured using Illumina techniques. For each type of mRNA, there is a specific secondary structure (SS). Using python scripts, SS sequences were divided in fragments and their frequencies were calculated. The initial dataset had the output variable as two possible classes (low or high mRNA expression) and 436 frequencies of different mRNA SS fragments (4067 cases). In the next step, a feature selection method from Weka software [4] was applied in order to obtain a reduced dataset (only 24 features). Machine Learning (ML) techniques from Weka were used to find the best classification model that can predict mRNA expression levels.

Results and Discussion: The final dataset of 24 selected features was the input of different ML techniques from Weka, such as LibLINEAR, BayesNet, NaiveBayes, MultilayerPerceptron, RandomForest. The best model is a NaiveBayes classifier with the true positive rate (TPR) for the low mRNA expression class of 0.78 (average true positive rate of 0.66). These results demonstrate the necessity for better models in future works, with different types of ML technique and other sets of mRNA SS features.

References:

1. Murrieta, C.M.; Hess, B.W.; Scholljegerdes, E.J.; Engle, T.E.; Hossner, K.L.; Moss, G.E.; Rule, D.C. Evaluation of milk somatic cells as a source of mrna for study of lipogenesis in the mammary gland of lactating beef cows supplemented with dietary high-linoleate safflower seeds. *J. Anim. Sci.* **2006**, *84*, 2399-2405.
2. Ma, J.L.; Zhu, Y.H.; Zhang, L.; Zhuge, Z.Y.; Liu, P.Q.; Yan, X.D.; Gao, H.S.; Wang, J.F. Serum concentration and mrna expression in milk somatic cells of toll-like receptor 2, toll-like receptor 4, and cytokines in dairy cows following intramammary inoculation with escherichia coli. *J. Dairy Sci.* **2011**, *94*, 5903-5912.
3. Witten, I.; Frank, E. *Data mining: Practical machine learning tools and techniques, second edition (morgan kaufmann series in data management systems)*. Morgan Kaufmann: 2005.
4. Smith, T.C.; Frank, E. Introducing machine learning concepts with weka. In *Statistical genomics: Methods and protocols*, Springer: New York, NY, 2016; pp 353-378.