**Computational Characterization of Exogenous MicroRNAs That Can Be Transferred into Human Circulation**

Jiang Shu1, Kevin Chiang1, Janos Zempleni2, Juan Cui1,\*

1Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE, USA

2Department of Nutrition and Health Sciences, University of Nebraska-Lincoln, Lincoln, NE, USA

\*To whom correspondence should be addressed.

Corresponding Author:

Juan Cui; jcui@unl.edu; 122B Avery Hall, Lincoln, NE 68588, USA

**Supplementary Methods**

**Addressing Data Imbalance Issue in Classification**

To overcome the imbalance problem that presents challenges in SVM-based classification [1], synthetic minority over-sampling technique (SMOTE) [2] was utilized to produce a balanced dataset for each kingdom separation. The main advantage of SMOTE when compared to other methods is that it enlarges the minority class region while making the minority sample less specific. Specifically, by using SMOTE, the production of new samples was not dependent on the few existing instances, but the *k* nearest neighbors. In addition, the discrete features and continuous feature are addressed individually in SMOTE:

* For each discrete feature:
	+ Finding majority vote *A* between one existing minority sample and its *k* nearest minority class neighbors based on the value distance metrics;
	+ Set the corresponding discrete feature as *A* in a new synthetic minority class sample.
* For continuous features:
	+ Calculate the difference *d* between a feature vector *v* of one existing minority sample and its *k* nearest minority class neighbors based on the Euclidean distance;
	+ Randomly generate a number *m* in [0,1] and calculate *d’*=*md*;
	+ Set the corresponding feature vector of a new synthetic minority class sample as *v’ = v* + *d’*.

By using SMOTE, for each kingdom separation, we generated numbers of minority class instances to improve the performance of classification.

**Classification-based feature selection**

As described above, by using SMOTE, we generated a number of synthetic minority class samples to produce a balanced dataset for kingdom separations. Then, a support vector machine (SVM)-based feature elimination strategy was introduced to find a set of most discriminative features for each classification. The processes of each kingdom separation were:

1. For a balanced dataset, we calculated F-score [3] of each features (1102, in total) with all samples;
2. According to the F-score ranking of all remained features (initial: 1102), we partitioned the entire dataset into 11 subsets.

For example, we generated the following subsets in the first round with all 1102 features:

* + SET1: all 1102 features;
	+ SET2: top 1000 features;
	+ SET3: top 900 features;

…

* + SET11: top 100 features.
1. For each subset, we conducted 5-fold cross validation. For each classification, we used 4 parts of data to training SVM model [3] and predicted on the rest 1 part. Repeated this process 5 times for each subset. Then, we evaluated overall classification performance by sensitivity, specificity, accuracy, and Matthews correlation coefficient (MCC):









1. Select the subset with the best classification performance, and repeated step II and III, until we found the minimal set of most discriminative features.

We re-estimated the parameters for each SVM-training and testing to ensure optimized model was found for each classification.

**The manifold ranking and parameter setting**

1. Calculate the distance *d* between any pair of samples in the entire dataset (both positive and background) and increasingly sort *d* of all pairs;
2. Construct a graph: represent each sample as a vertex in a graph, and repeat connecting two vertices according to sorted distance list until the whole graph is connected;
3. Form a weight matrix *W*:



Since there is no self-loop in the graph, thus ;

1. Form a diagonal matrix *D* that . Symmetrically normalize *W* by



1. Iterate the ranking score function for each sample until  converges:



1. Use  denotes the final ranking score of each samples, and decreasingly sort .

For parameter , we used the empirical setting, ****, as discussed in [4, 5]. In the ranking score function,  is the index of input, where "1" denotes the sample is from positive class and "0" means the sample is from the background class;  decides how much weight we applied to the prior knowledge when calculating the ranking score for each sample. In order to minimize the prior knowledge from the input, we fixed  as 0.01 in this study.

By the end of Manifold ranking, we expect a final ranking of all miRNAs that are included in this study. Different from the typical binary classification that labels “yes or no” on each entry, the ranking strategy indicates the relative transportability of each miRNA into circulation. Without an explicit threshold for transportability, we can infer the highly possible candidates by examining their ranks with respect to known human blood miRNAs. Specifically, the miRNAs ranked above any known circulating miRNAs are likely human absorbable miRNAs.

In particular, to avoid over-fitting problem, our method integrates the following processes: An ensemble SVM model is designed to identify a set of most discriminative features. The five-fold cross validation was conducted to assess the performance on the selected features and the randomly-generated training/testing datasets. Moreover, for each SVM training and testing process, the model employed a parameter optimization process that ensures the best performance was achieved. A hybrid feature set is used for the final manifold ranking, which include not only the features that are specific to human blood circulating miRNA classification (360 human blood miRNAs against other human miRNAs), but also the discriminative features in each kingdom-wise separation.

**Intestinal transport of milk Extracellular Vesicles (EVs) in human cells**

Exosomes were purified from cow’s milk using ultracentrifugation [6] and labeled with the fluorophore, FM4-64 to allow for tracking of uptake into fully differentiated human intestinal Caco-2 cells [7]. Culture medium was replaced with human EV-depleted medium that was supplemented with fluorophore-labeled exosomes to produce a concentration of 600fmol/L miR-29b. Uptake of EVs, and therefore miRNAs, into Caco-2 cells was linear with time for up to 2 hours. The uptake of EVs is a carrier-mediated[8] process, based on the following lines of evidence. 1) The uptake exhibited saturation kinetics that could be modeled using the Michaelis-Menten equation. 2) The uptake of EVs was 80% lower when studies were conducted at 4°C or when surface proteins were removed from EVs or cells by proteinase K treatment, compared with untreated controls at 37°C. Note that surface proteins are known to play essential roles in the uptake of human EVs in human cells [8]. The existence of intestinal EV/miRNA carriers supports our theory that dietary miRNAs might be essential for maintaining optimal health, and that postprandial increases in plasma miRNAs are caused by absorption of dietary miRNAs rather than an increase in the expression of genes coding for miRNAs.

**Experimental detection of dietary miRNAs in human circulation**

Plasma was collected from a milk-feeding study [9] where five healthy adults consume three doses of 1% fat milk (0.25L, 0.5L and 1.0L) and an ongoing egg feeding study. The miRNA abundance in plasma was quantified by qRT-PCR using a miScript SYBR Green PCR Kit while MiR-1 was used as a negative control because its levels in milk are below current detection limits. Details about the intestinal transport test of milk EVs in human cells are given in Supplementary Methods. MiR-29b and miR-200c are found with increased expression in subjects in cow-milk feeding study and one egg miRNA, gga-miR-1451, has been identified in human circulation.

**Supplementary Reference:**

1. J. Brank MG, N. Milić-frayling , D. Mladenić. Feature selection using support vector machines. In Proc of the 3rd Int Conf on Data Mining Methods and Databases for Engineering, Finance, and Other Fields. 2002.

2. Nitesh V. Chawla KWB, Lawrence O. Hall , W. Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002.

3. Chih-Chung Chang C-JL. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011.

4. Liu Q, Cui J, Yang Q, Xu Y. In-silico prediction of blood-secretory human proteins using a ranking algorithm. BMC bioinformatics. 2010;11:250. doi: 10.1186/1471-2105-11-250. PubMed PMID: 20465853; PubMed Central PMCID: PMC2877692.

5. Zhao YF, He LY, Liu BY, Li J, Li FY, Huo RL, et al. Syndrome classification based on manifold ranking for viral hepatitis. Chinese journal of integrative medicine. 2014;20(5):394-9. doi: 10.1007/s11655-013-1659-4. PubMed PMID: 24174345.

6. Hata T, Murakami K, Nakatani H, Yamamoto Y, Matsuda T, Aoki N. Isolation of bovine milk-derived microvesicles carrying mRNAs and microRNAs. Biochemical and biophysical research communications. 2010;396(2):528-33. doi: 10.1016/j.bbrc.2010.04.135. PubMed PMID: 20434431.

7. Tian T, Wang Y, Wang H, Zhu Z, Xiao Z. Visualizing of the cellular uptake and intracellular trafficking of exosomes by live-cell microscopy. Journal of cellular biochemistry. 2010;111(2):488-96. doi: 10.1002/jcb.22733. PubMed PMID: 20533300.

8. Escrevente C, Keller S, Altevogt P, Costa J. Interaction and uptake of exosomes by ovarian cancer cells. BMC cancer. 2011;11:108. doi: 10.1186/1471-2407-11-108. PubMed PMID: 21439085; PubMed Central PMCID: PMC3072949.

9. Baier SR, Nguyen C, Xie F, Wood JR, Zempleni J. MicroRNAs are absorbed in biologically meaningful amounts from nutritionally relevant doses of cow milk and affect gene expression in peripheral blood mononuclear cells, HEK-293 kidney cell cultures, and mouse livers. The Journal of nutrition. 2014;144(10):1495-500. doi: 10.3945/jn.114.196436. PubMed PMID: 25122645; PubMed Central PMCID: PMC4162473.