Feature Extraction and Discovery of microRNAs Using Nonnegative Matrix Factorization

Weixiang Liu[1] **Tianfu Wang**[1] **Siping Chen**[1] **Aifa Tang**[2]

[1]Shenzhen Key Lab of Biomedical Engineering
Shenzhen University, Shenzhen, China 518060
[2]Shenzhen Key Lab of Male Reproduction and Genetics
Peking University Shenzhen Hospital, Shenzhen, China 518036

Abstract

In this paper we consider a computational approach for feature extraction and discovery of microRNA (miRNA) sequences based on the hairpin precursors. We firstly computed the frequency of one nucleotide or one sub-sequence combined from several nucleotides in each hairpin precursor miRNA. Then we compared nonnegative matrix factorization and principal component analysis methods to decompose the frequency matrix for further analysis, such as miRNA gene discovery. Our experimental results on recently published miRNA sequences and the same number of random sequences demonstrate the effectiveness of the proposed feature extraction method, and especially that nonnegative matrix factorization significantly outperforms principal component analysis in miRNA gene discovery.

Keywords: microRNA, nonnegative matrix factorization, principal component analysis, gene discovery, feature extraction

1. Introduction

In genetics, the mature microRNAs (miRNA) are single-stranded RNA molecules of about 21-23 nucleotides in length thought to regulate the expression of other genes [1]. Over the last five years, miRNA related topics and research projects grow fast. For example, the number of published miRNA sequences increases 20 folds from 2002 to 2006 [2]; and up to May 2007, 4584 miRNA genes have been reported in the current miRBase sequence database (release version 9.2) [1].

Discovery of miRNA genes is one of the most imminent problems towards understanding of post-transcriptional gene regulation. There two kinds of methods to this end: experimental cloning and computational prediction. Experimental cloning efforts have successfully identified highly expressed miRNAs from various tissues. However computational prediction of miRNAs should be a robust approach for tissue-specific or lowly expressed miRNAs. Several computational methods have been developed to find close homologs among related miRNAs [3]. These methods can be divided into three groups:

- Using secondary structure information;
- Relying on phylogenetic conservation of both sequence and structure;
- Assessing the thermodynamic stabil-

[1] http://microrna.sanger.ac.uk/registry/

Proceedings of the 11th Joint Conference on Information Sciences (2008)
Published by Atlantis Press

ity of hairpins and sequence and structure or using information on genomic location.

In this paper we use the information of frequency of one or more nucleotides in each sequence. In fact, all miRNA gene sequences consist of 4 nucleotides: A, U, C and G. Considering words from 26 alphabets in a document, we can adopt language models or text/document analysis methods for biological sequences; e.g. see [4, 5]. We follow this way and apply a novel machine learning method, nonnegative matrix factorization (NMF) [6], to miRNA gene sequence analysis and discovery. Our method contains two important steps:

1. Calculate frequency matrix of one or nucleotides in each sequence;
2. Decompose the matrix using NMF.

Finally we can use the reduced data via NMF for further analysis, such as visualization or miRNA gene discovery.

The rest of this paper is organized as below. In section 2 we introduce our feature extraction method and discuss NMF for feature extraction and clustering analysis. Experimental results are shown in Section 3 and we conclude in Section 4.

2. Methods

2.1. Feature extraction

Numerous studies have shown that oligonucleotide frequencies within DNA sequences is a promising measure to sequence analysis [7, 8]. Our feature extraction method is based on such measure. Given a miRNA sequence $\mathfrak{S}$ and a string str containing A, or AU, or AUC, or AUCG, ..., we can calculate the frequency of str in $\mathfrak{S}$, denoted as $f_s(str)$ for simplicity, where s is the length of str. In this paper we consider all combinations of one or more nucleotides for analysis. For a sequence, we define the feature vector $f_s, s = 1, 2, 3, \ldots$, as below

$$\begin{aligned} f_1 &= [f(A)\ldots f(G)] \in R_+^4, \\ f_2 &= [f(A)\ldots f(GG)] \in R_+^{20}, \\ f_3 &= [f(A)\ldots f(GGG)] \in R_+^{84}, \\ &\ldots \end{aligned}$$

where R_+ means the element of feature vector f_i is nonnegative. So given m sequences and considering the case of $s = 3$, we can get the frequency matrix as $X \in R_+^{84\times m}$. Generally speaking, the more nucleotides are chosen, the larger the length of the feature is. In our following experiments, we compared four cases, i.e. $s = 1, 2, 3$ and 4.

2.2. Nonnegative matrix factorization for feature extraction and clustering

NMF is a new method for nonnegative data analysis. The method considers nonnegative constraint on matrix factorization and has some advantages over traditional PCA [6], especially in face image analysis and document analysis. Now it becomes a powerful technique for nonnegative data analysis; see [9, 10, 11] for recent literature review and more references therein.

Given a data matrix $\mathbf{X}$ with nonnegative values for all entries, NMF is to find an approximation decomposition $X \approx WH$, where W is $n \times k$ and H is $k \times m$, with nonnegative constraint on both W and H. There are two algorithms with multiplicative updates for NMF as [12]

$$W_{ik} \leftarrow W_{ik} \frac{(XH^T)_{ik}}{(WHH^T)_{ik}},$$

$$H_{kj} \leftarrow H_{kj} \frac{(W^TX)_{kj}}{(W^TWH)_{kj}},$$

and

$$W_{ik} \leftarrow W_{ik} \frac{\sum_j H_{kj} X_{ij} / (WH)_{ij}}{\sum_j H_{kj}},$$
$$H_{kj} \leftarrow H_{kj} \frac{\sum_i W_{ik} X_{ij} / (WH)_{ij}}{\sum_i W_{ik}}.$$

These multiplication update rules can hold nonnegativity easily with nonnegative initialization. In our following experiments, we adopt the basic rule as in paper [6].

Given a factorization as above, setting $k < n$ leads to dimension reduction. And in this sense, NMF is similar to PCA. As discussed in [6], NMF is also similar to K-means clustering method. Recently Brunet et al. applied NMF for clustering gene expression data [13]. The clustering rule is defined as below according to the values of H: assign sample j to cluster k if the H_{kj} is the largest element in column j [13]. In [14] NMF is used to cluster genes.

In K-means method, each sample can only be assigned to one class. This is also called "hard" clustering in which the class membership of one sample is 1 or zero. By contrast, in "soft" clustering, the class membership of each sample belongs to a value in $[0,1]$. Fuzzy clustering and probabilistic clustering are two popular "soft" clustering methods. In this sense, NMF is "soft" and K-means "hard". From the view point of matrix factorization, in K-means or vector quantization algorithm, each column of H is a unary vector with only one element being one and other elements being zeros [6]. For clustering, both NMF and K-means are stochastic, but NMF does not depends on the nearest neighbor criterion of K-means. More recently, the equivalence of nonnegative matrix factorization and spectral clustering is discussed in detail [15].

For real applications, Lee and Seung discussed two special cases: gray face images and semantic features of documents [6]. And this demonstrates that NMF can be used directly on nonnegative data, e.g. gray images, and it also can be used on transformed data with nonnegativity, e.g. frequency matrix of words in documents. In [9], NMF has been used for system call based intrusion detection, gray image watermarking, and EEG analysis. In [10, 11] more references on applications of NMF are available. And in this study we apply NMF for miRNA sequence analysis and discovery.

3. Experimental results

We tested our method on two datasets. One is the miRBase sequence database with release version 9.2 [2]. It contains 4584 entries representing hairpin precursor miRNAs, expressing 4430 mature miRNA products, in primates, rodents, birds, fish, worms, flies, plants and viruses. The data are freely available to all through the web interface at http://microrna.sanger.ac.uk/sequences/. The other is 4584 random sequences with same length corresponding to each of the miRBase sequence database.

We firstly applied NMF for reducing the data with case of $s = 1$ for 2-D visualization as shown in Figure 1. And we also compared NMF and PCA on the data. We can see that: 1) for representation, our feature extraction method is effective because microRNAs and random sequences can be divided separately to some extent while 2) the features reduced by NMF are more discriminative for two groups than those by PCA.

Now we used NMF for discovering miRNAs with the discussed clustering rule and we also compared NMF and PCA. For PCA, the coefficients projected on principal components are all changed to positive with absolute function. Here we considered the data with four cases: $s = 1,2,3$ and 4. For NMF, we calculated the average of clustering accuracies over 100 runs

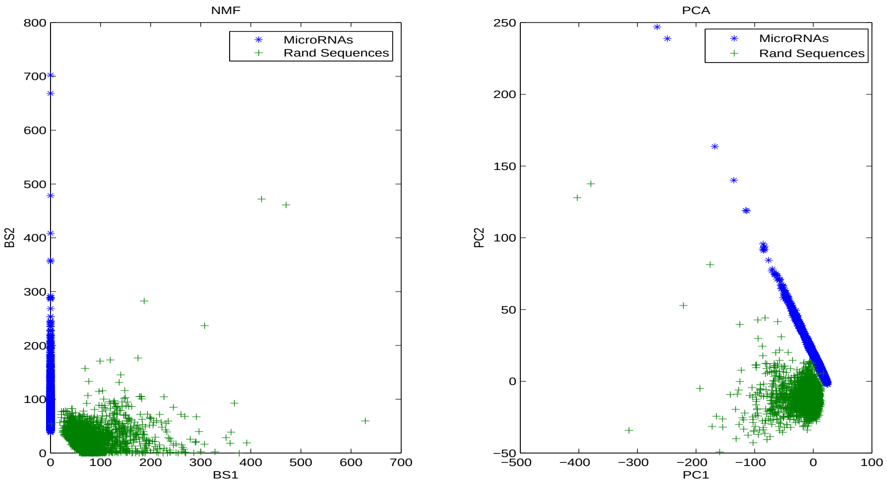

Fig. 1: 2-D visualization of microRNAs and random sequences by NMF and PCA where "BS" standing basis sequence and "PC" for principal component.

Table 1: miRNA discovery results (%) on true 4584 sequences and 4584 random sequences with NMF and PCA.

Case	$s = 1$	$s = 2$	$s = 3$	$s = 4$
NMF	94.6	97.8	98.7	99.1
PCA	65.4	66.9	67.4	67.4

as the final result. As shown in Table 1, we can see NMF significantly outperforms PCA. In addition, we can see that the accuracy was improved with more features considered and the reason my be that a string with many nucleotides can capture more structure information from the sequence than one nucleotide.

4. Conclusions and future work

We proposed a computational approach for feature extraction and discovery of microRNA (miRNA) sequences using nonnegative matrix factorization with hairpin precursor miRNAs. We firstly computed the frequency of one or more nucleotides in each sequence as a feature of the sequence. Then we compared nonnegative matrix factorization and principal component analysis methods to decompose the frequency matrix for further analysis, such as visualization and miRNA gene discovery. Our primary experimental results on recently published 4584 microRNA sequences and the same number of random sequences demonstrate the effectiveness of the proposed feature extraction method, and especially that nonnegative matrix factorization outperforms principal component analysis in miRNA gene discovery.

Currently we are investigating the proposed method for novel microRNA gene identification and verify it with biology experiments as future work. In addition, our feature extraction method as used in document analysis is also useful for other biological sequence analysis, such as sequence alignment and sequence similarity investigation.

Acknowledgements

This work was supported by the Open Fund of Shenzhen Key Lab of Biomedical Engineering and the National Natural Science Foundation of China (No. 30700824). The authors would like to thank all researchers who made the data publicly available.

References

[1] G. Ruvkun. Molecular biology: Glimpses of a tiny RNA world. *Science*, 294(5543):797–799, 2001.

[2] S. Griffiths-Jones, R.J. Grocock, S. van Dongen, A. Bateman, and A.J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(Database Issue):D140–D144, 2006.

[3] E. Berezikov, E. Cuppen, and RH Plasterk. Approaches to microRNA discovery. *Nat. Genet*, 38(Suppl 1):S2–S7, 2006.

[4] D.B. Searls. The language of genes. *Nature*, 420(14):211–217, 2002.

[5] R. McMahon. Genes and Languages. *Community Genetics*, 7(1):2–13, 2004.

[6] D D Lee and H S Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

[7] N. Goldman. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acids Res*, 21(10):2487–91, 1993.

[8] D.T. Pride, R.J. Meinersmann, T.M. Wassenaar, and M.J. Blaser. Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases. *Genome Research*, 13(2):145–158, 2003.

[9] W.X. Liu, N.N. Zheng, and Q.B. You. Nonnegative Matrix Factorization and its applications in pattern recognition. *Chinese Science Bulletin*, 51(17-18):7–18, 2006.

[10] M. Berry, M. Browne, A. Langville, P. Pauca, and R. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. Submitted to Computational Statistics and Data Analysis, 2006.

[11] S. Sra and I.S. Dhillon. Nonnegative matrix approximation: Algorithms and applications. Technical Report TR-06-27, Department of Computer Sciences, The University of Texas at Austin, 2006.

[12] D D Lee and H S Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13*, pages 556–562, 2001.

[13] J P Brunet, P Tamayo, T R Golub, and J P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA*, 101(12):4164–4169, 2004.

[14] PM Kim and B Tidor. Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res*, 13:1706–1718, 2003.

[15] C. Ding, X. F. He, and H. D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proc. SIAM Int'l Conf. Data Mining (SDM'05)*, April 2005.