Cancer classification using entropy analysis in fractional Fourier domain of gene expression profile

ZhiPeng Guo, Yi Xin and YiZhang Zhao

Department of Biomedical Engineering, School of Life Science, Beijing Institute of Technology, Beijing, P.R. China

ABSTRACT
The vast advancement in the field of DNA microarrays has enabled researchers to simultaneously analyse the expression levels of thousands of genes on a single microarray chip. Several data-mining methods have been applied in studying the gene expression profiles to distinguish between various sub-types of cancer and types of other diseases. However, accurate diagnosis of cancer sub-types remains a challenge. The gene-by-gene-based approaches are likely to produce chance correlations owing to the high-dimensional nature of the microarray experiments, and the fact that biological phenomena are constantly cyclical and rhythmic. Gene expression is highly regulated and correlations exist between the expressions of different genes; therefore, the cooperativity of genes must be taken into consideration to capture the inherent characteristics of the genome. The present method utilized fractional Fourier transform (FRFT) and entropy-based techniques to extract representative features of the gene expression profile (GEP) and conduct tumour classification using the support vector machine (SVM). The algorithm was tested using four different data-sets. The experimental results reveal that this algorithm has the ability to classify cancers into various types and sub-types with high accuracy.

Introduction
Classification of microarray gene expression data has become an important topic of research in bioinformatics owing to its potential in aiding medical diagnosis. An increasing number of new classification, prediction and clustering techniques are being used for the analysis of data based on the nature of small-sized samples and the high-dimensional feature of gene expression data [1–4]. In 2015, Nguyen et al. [5] used Hidden Markov models to classify a diffuse large B-cell lymphoma data-set (DLBCL), leukaemia data-set, colon data-set and prostate data-set. The rate of accuracy of these tests was found to be 98.83%, 98.26%, 89.11% and 92.01%, respectively [5]. In 2016, Ghadiri et al. [6] used a method that combined linear discriminant analysis (LDL) and partial least squares (PLS) to classify lung cancer, and the rate of accuracy obtained was over 94.5%. Most of the methods in tumour classification research, such as the gene-by-gene-based approach, ignore the cooperativity of genes [7]. However, gene expression is a highly regulated phenomenon; therefore, it is important to acquire significant genetic characteristics on a genome-wide scale.

This study combined the fractional Fourier transform (FRFT) and entropy-based technique to analyse the gene expression profile (GEP). First, the raw data were transformed into a special fractional Fourier domain with a selected order of FRFT, which had been verified for suitable the pattern recognitions of biological signals as well as for reducing noise [8]. Next, FRFT was combined with an entropy-based method, which could extract inherent genetic features on a genome-wide scale. Finally, tumour classification was performed using the support vector machine (SVM). This method offers a number of advantages; for example, this algorithm has the ability to classify cancers into various sub-types with high accuracy and reflect the inherent relativity of gene networks.

Materials and methods
GEP can be treated as a set of non-stationary physiological signals, referring to the inherent structure of genomes in cancerous or healthy individuals [9]. Mathematically, let $X = (x_1, x_2, \ldots, x_n)$ be a set of genes and $G = \{g_1, g_2, \ldots, g_n\}$ be a set of samples. The corresponding gene expression matrix named GEP can be expressed as $X = (x_{ij})_{m \times n}$, where each row represents a sample for tumour diagnosis and each column represents a gene from a different sample. The numerical
value of $x_{ij}$ denotes the expression level of a specific gene $j (j = 1, 2, \ldots, n)$ of a particular sample $i (i = 1, 2, \ldots, m)$ [10]. On a physiological standpoint, there exists a wide cooperation between genes; in addition, a vast number of excess genes also exist, which are not related to other genes. In other words, there are messy ‘noises’ in the form of signals in the GEP [11]. Obviously, FRFT can analyse a signal on a genome-wide scale, which has been demonstrated to be more appropriate for feature extraction and noise suppression in non-stationary signal processing, specifically in case of physiological signals [8].

**Fractional Fourier transform**

The fractional Fourier transform (FRFT) is a generalized form of the ordinary Fourier transform [8]. FRFT is a linear operator as defined previously [12–15].

$$f_p(u) = \int_{-\infty}^{+\infty} f(t)K_p(t, u)\, dt$$  \hspace{1cm} (1)

With $K_p(t, u)$ representing the kernel function defined as

$$K_p(t, u) = \begin{cases} A_u \exp[j\pi (u^2 \cot \alpha - 2ut \csc \alpha + t^2 \cot \alpha)], \alpha \neq n\pi \\ \delta(u - t), \alpha = 2n\pi \\ \delta(u + t), \alpha = (2n \pm 1)\pi \end{cases}$$

and $A_u = \exp[-j\pi \text{sgn}(\sin \alpha)/\alpha + j\alpha/2], \alpha = \frac{\alpha\pi}{2}, n$ is an integer; $\delta(t)$ represents the Dirac function. For $\alpha = 2\pi + \pi/2$, the FRFT translates into the conventional Fourier transform. Any intermediate value of $\alpha (0 < \alpha < \pi/2)$ produces a rotated time-frequency representation of the signal [16,17].

**Entropy-based method**

In thermodynamics, entropy [18] is a measure of disorder. The entropy method is a kind of objective weighting method; it calculates the weight of indexes by entropy.

Subjective fixed-weight methods such as the Delphi method are usually used when determining weights of indexes [19,20]. Such methods could lead to subjective deviations. The entropy weight method is an objective fixed-weight method, which utilizes the quantity of information to determine the weight of an index of interest. Such methods are based on the nature of indexes to determine their weights, which could eliminate subjective deviations and ensure that the results are more concurrent with the facts. In this study, the entropy weight method was used to calculate the weight of the amplitude of the FRFT coefficient, which is calculated as follows [21].

First, assuming there are $i$ samples and $j$ FRFT coefficients, $x_{ij}$ is the $j$th amplitude of FRFT coefficient in the $i$th sample. In order to eliminate the influence of the coefficient, it is essential to normalize the value of the gene by relative optimum membership degree, because the index in different dimensions is incommensurable. To the benefit of the FRFT coefficient, the attribute value of the $j$th FRFT coefficient in the $i$th sample can be transformed by Equation (2).

$$t_{ij}' = \frac{x_{ij}}{\max x_{ij}}, (i = 1, \ldots, m; j = 1, \ldots, n)$$  \hspace{1cm} (2)

To the cost FRFT coefficient, the attribute value of the $j$th FRFT coefficient in the $i$th sample can be transformed by Equation (3).

$$t_{ij}' = \frac{\min x_{ij}}{x_{ij}}, \min x_{ij} \neq 0, (i = 1, \ldots, m; j = 1, \ldots, n)$$  \hspace{1cm} (3)

After normalization of the value of the FRFT coefficient, the standardized FRFT coefficient matrix can be obtained as $R = [r_{ij}]_{m \times n}$.

Second, the entropy of the $j$th FRFT coefficient is determined by Equation (4) according to the definition

$$H_j = -\sum_{i=1}^{m} \frac{f_{ij} \ln f_{ij}}{\ln m}, (i = 1, \ldots, m; j = 1, \ldots, n)$$  \hspace{1cm} (4)

$$f_{ij} = \frac{t_{ij}'}{\sum_{j=1}^{m} t_{ij}'}, (i = 1, \ldots, m; j = 1, \ldots, n)$$  \hspace{1cm} (5)

Finally, the entropy weight of the $j$th FRFT coefficient is determined by Equation (6).

$$w_j = \frac{1 - H_j}{n - \sum_{j=1}^{n} H_j}, \sum_{j=1}^{n} w_j = 1, (j = 1, \ldots, n)$$  \hspace{1cm} (6)

The entropy weight reflects the quantity of useful information of the FRFT coefficient in information theory. Therefore, the higher the entropy weight of the FRFT coefficient is, the more useful the information of the FRFT coefficient is, and vice versa.

**Description of the tumour classification algorithm**

The algorithm model is represented as follows:

Step 1. Each row of the GEPX ($i = 1, 2, \ldots, m$) is transformed into an FRFT domain as $x^i (i = 1, 2, \ldots, m)$; each row represents the genes of one sample.
Step 2. Because the global-scale features could be accurately correlated with the whole genome, the amplitude values of \( x^a_i \) \((i = 1, 2, \ldots, m)\) are treated as a new data-set \( a_i \) \((i = 1, 2, \ldots, m)\) for feature selection with the entropy method.

Step 3. The value of the FRFT coefficient is normalized by Equation (2).

Step 4. The entropy of the FRFT coefficient is calculated by Equation (4).

Step 5. The entropy weight of the FRFT coefficient is calculated by Equation (6).

Step 6. The 300 top-ranked FRFT coefficients set by the entropy weight are chosen.

Step 7. The new matrix of 300 FRFT coefficients obtained from the entropy-based method is classified using the four-fold cross-validation method by the SVM [5,22–25].

Step 8. Steps 1–7 are repeated by varying the parameter \( \alpha \) until the best result is found.

Results and discussion

Data-sets

Four public microarray data-sets were used to test this algorithm.

Leukaemia data

The leukemia data-set, obtained from the Massachusetts Institute of Technology (MIT) database, contained 7129 gene expression levels of 72 samples. Two variants of leukemia existed in the samples: AML (Acute Myeloid Leukemia), 25 samples; and ALL (Acute Lymphoblastic Leukemia), 47 samples.

Colon cancer data

This data-set, obtained from the Princeton University database, contained 22 normal and 40 tumorous colon tissue samples. Only 2000 genes of the highest minimal intensity were selected [26]. The data were pre-processed by logarithmic computation and standardization.

Gastric cancer data

This data-set, obtained from the CODEBUS database, contained 20 normal and 20 gastric cancer tissue samples. Only 1519 genes with the highest minimal intensity were selected.

Breast cancer data

This data-set, obtained from the University of California at Irvine (UCI) database, contained 45 normal and 41 breast cancer tissue samples. Only 7129 genes of the highest minimal intensity were selected [27].

Experimental results

In the experiments, the optimal order of FRFT varied from 0 to 1. Figures 1 and 2 show the performance of the optimal order of FRFT in the range from 0 to 1 with a step of 0.05. The four-fold cross-validation method was employed to make the results more credible. Figure 1 indicates that the cross-validation accuracy of the leukemia data-set is the best when the optimal order is 0.1, 0.65 and 0.95; the cross-validation accuracy of the gastric data-set is the best when the optimal order is 0.35; the cross-validation accuracy of the colon data-set is the best when the optimal order is 0.8; and the cross validation accuracy of the breast data-set is the best when the optimal order is 0.05, 0.2, 0.45 and 0.5.

Next, the best accuracy was sorted into Table 1 as follows.

By setting the optimal order of FRFT to the best result, the sensitivity, specificity, accuracy and Youden’s index was calculated. Table 2 shows the sensitivity, specificity, accuracy, and Youden’s index of these four data-sets.
which confirmed that the result of this algorithm is credible and has a high accuracy.

Compared with other studies, this method can recognize different types of gene spectrum data more quickly and maintain a high accuracy. Moreover, the results indicate that genes function as a network is connected to the multistability feature of tumour tissues [28]. The estimation of an optimal order for each tumour data-set is still at an experimental stage, owing to lack of prior knowledge from certain strict mathematical theories. However, the optimal order might be a significant potential indicator for measuring the correlations among genes, which could be validated from the distinct best orders, as observed in the 4 validation sets of our study.

Conclusions

A method for tumour detection and classification, mainly based on FRFT and entropy, was proposed in this paper. FRFT is more useful in correlation and collaborative analysis of data. This method extracts the feature of data in whole sequence not, not in a single part or segment. The multistability feature can be extracted. The experimental results proved that this method could achieve high accuracy in tumour classification.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Granzow M, Berrar D, Dubitzky W, et al. Tumor classification by gene expression profiling: comparison and validation of five clustering methods. Acm Sigbio Newsletter. 2001;21:16–22.

[2] Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: a survey. IEEE Comput Soc. 2004;11:1370–1386.

[3] Pham TD, Wells C, Crane DL. Analysis of microarray gene expression data. Curr Bioinform. 2006;1:37–53.

[4] Asyali MH, Colak D, Demirkaya O, et al. Gene expression profile classification: a review. Curr Bioinform. 2006;1:55–73.

[5] Nguyen T, Khosravi A, Creighton D, et al. Hidden Markov models for cancer classification using gene expression profiles. Inform Sci. 2015;316(C):293–307.

[6] Ghadiri K, Rezaei M, Tabatabaei SM, et al. Use of LDA combined with PLS for classification of lung cancer gene expression data. Int J Med Res Health Sci. 2016;5(9S):500–506.

[7] Wang SL, Li X, Zhang S, et al. Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction. Comput Biol Med. 2010;40(2):179–189.

[8] Ozaktas HM, Zalevsky Z, Kutay MA. The fractional Fourier transform: with applications in optics and signal processing. [Place unknown]: Wiley; 2001.

[9] Li T. A survey on wavelet applications in data mining. ACM SIGKDD Explor Newsletter. 1993;4(2):49–68.

[10] Zheng CH, Wu FL, Li B, et al. Constrained maximum variance mapping for tumor classification. Emerging intelligent computing technology and applications. Springer Berlin Heidelberg,2009:102-111.

[11] Golub TR, Slonin DK, Golub TR. Molecular classification of cancer. J Clin Microbiol. 1000;93(5):1210–1220.

[12] Mendlovic D, Ozaktas HM. Fractional Fourier transforms and their optical implementation: part I. J Optical Soc Am A: Optics Image Sci Vis. 1993;10(9):1875–1881.

[13] Mendlovic D, Ozaktas HM. Fractional Fourier transforms and their optical implementation: part II. J Optical Soc Am A: Optics Image Sci Vis. 1993;10(12):2522–2531.

[14] Larsson EM, Sundén P, Olsson CG, et al. Fractional Fourier transform and time-frequency representations. Am J Roentgenol. 2003;180(1):227–232.

[15] Cariolaro G, Ersege H, Kraniauskas P, et al. A unified framework for the fractional Fourier transform. IEEE Trans Signal Process. 1998;46(12):3206–3219.

[16] Sejdić E, Djurović I, Jiang J. Time-frequency feature representation using energy concentration: an overview of recent advances. Digital Signal Process. 2009;19(1):153–183.

[17] Sejdić E, Djurović I, Jiang J, et al. Time-frequency based feature extraction and classification: considering energy concentration as a feature using the stockwell transform and related approaches. [Place unknown]: VDM Verlag Dr. Müller; 2009.

[18] Jungermann AH. Entropy and the Shelf Model: a quantum physical approach to a physical property. J Chem Educ. 2006;83(11):1686–1694.

[19] Wang YF, Dai M. Safety assessment based on information entropy and unascertained theory. Shaanxi Coal. 2009;5:8–10.

[20] Chen ZY, Yang SQ, Wu CF, et al. Application of uncertainty measurement model based on coefficient of entropy to comprehensive evaluation of coal mine safety. Mining Safety Environ Protect. 2007;34(1):75–77.

[21] Li X, Wang K, Liu L, et al. Application of the entropy weight and topsis method in safety evaluation of coal mines. Procedia Eng. 2011;26(4):2085–2091.

[22] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46(1–3):389–422.
[23] Rakotomamonjy A. Variable selection using SVM based criteria. J Mach Learn Res. 2003;3(7–8):1357–1370.

[24] Weston BJ, Mukherjee S, Chapelle O, et al. Feature selection for SVMs neural information processing systems. J Aerosol Sci. 2013;28(Suppl 1):333–334.

[25] Zhang X, Wong W. Recursive sample classification and gene selection based on SVM: method and software description. Department of Biostatistics, Harvard School of Public Health; 2001. (Technical report).

[26] Alon BU, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide array. Proc Natl Acad Sci USA. 1999;96(12):6745–6750.

[27] West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA. 2001;98(20):11462–11467.

[28] Wilkins AS. For the biotechnology industry, the penny drops (at last): genes are not autonomous agents but function within networks. Bioessays. 2007;29(12):1179–1181.