

---

## **LIBGS: A MATLAB software package for gene selection**

---

Yi Zhang, Dingding Wang and Tao Li\*

School of Computing and Information Sciences,  
Florida International University,  
11200 SW 8th St., Miami, FL 33199, USA  
E-mail: yzhan004@cs.fiu.edu  
E-mail: dwang003@cs.fiu.edu  
E-mail: taoli@cs.fiu.edu  
\*Corresponding author

**Abstract:** Many gene selection algorithms have been applied in microarray expression data analysis successfully. To solve different developing environments of these toolkits, perform data analysis and algorithm comparison more flexible, we have developed a software package LIBGS including: rankgene (Su et al., 2003), and mRMR (<http://research.janelia.org/peng/proj/mrmr/index.htm>). Due to different developing environments of these toolkits, it is difficult to compare different algorithms using them. We have developed a software package named **LIBGS**, which can effectively evaluate gene function in discriminating biological samples of different types. This package includes:

- seven new gene selection algorithms implemented using MATLAB
- a MATLAB interface for Rankgene which includes another eight selection measures
- a MATLAB interface for two well-known classification tools (e.g., LIBSVM and WEKA)
- programs for converting data formats
- a collection of six popular gene expression data sets.

These features make LIBGS a useful tool in gene expression analysis and feature selection.

**Keywords:** LIBGS; gene selection algorithms; MATLAB.

**Reference** to this paper should be made as follows: Zhang, Y., Wang, D. and Li, T. (2010) 'LIBGS: A MATLAB software package for gene selection', *Int. J. Data Mining and Bioinformatics, Vol.*

**Biographical notes:** Yi Zhang received the BS Degree in the School of Telecommunications Engineering from Hangzhou Dianzi University, China. She received her Master's Degree in the School of Electrical, Computer and Telecommunications Engineering from University of Wollongong, Australia. Now, she is a PhD student in the School of Computing and Information Sciences, Florida International University. She works on the gene selection algorithms and text clustering.

Author: Please  
reduce abstract  
of no more than  
100 words.

Dingding Wang received her BS Degree in Computer Science from University of Science and Technology of China. Currently, she is a PhD student in School of Computing and Information Sciences, Florida International University.

Tao Li is currently an Assistant Professor in the School of Computing and Information Science, Florida International University. He received his PhD in Computer Science from the Department of Computer Science, University of Rochester in 2004. His research interests are in data mining, machine learning, information retrieval, and bioinformatics.

---

## 1 Introduction

The recent development of microarray technologies has enabled biologists to quantify gene expression of tens of thousands in a single experiment. One of the urgent yet fundamental issues is to identify a set of genes and its expression patterns that either characterise a certain cell state or predict a certain cell state in the future. Gene selection aims at finding a set of genes that best discriminate biological samples of different types. The selected genes are ‘biomarkers’, and they form ‘marker panel’ for analysis.

### 1.1 Gene selection algorithms

Many gene selection algorithms have been applied in gene expression data analysis successfully (Li et al., 2004; Marko and Igor, 2003; Peng et al., 2005; Ye et al., 2004; Zhang et al., 2007; Zheng, 2007; Zhu et al., 2008). In LIBGS, we provide 15 different gene selection methods for quantifying a gene’s ability to distinguish between classes. These different measures are briefly described as follows.

- *ReliefF*: ReliefF (Marko and Igor, 2003) is used to estimate the quality of genes according to how well their values distinguish between instances that are near to each other. Given a randomly selected instance  $I_m$  from class  $L$ , ReliefF searches for  $K$  of its nearest neighbours from the same class called nearest hits  $H$ , and also  $K$  nearest neighbours from each of the different classes, called nearest misses  $M$ . It then updates the quality estimation  $W_i$  for gene  $i$  based on their values for  $I_m$ ,  $H$ , and  $M$ . If instance  $I_m$  and those in  $H$  have different values on gene  $i$ , then the quality estimation  $W_i$  is decreased. On the other hand, if instance  $I_m$  and those in  $M$  have different values on the the gene  $i$ , then  $W_i$  is increased.
- *mRMR*: mRMR selects genes that have the highest relevance with the target class and are maximally dissimilar to each other (Peng et al., 2005). For discrete/categorical variables, MID and MIQ which are mutual information based solutions are widely used to measure the level of ‘similarity’ between the genes and the discriminant powers of genes with different target class. In the case of continuous variable, FCD and FCQ are effective solutions, which use the F-statistic between the genes and the class to score maximum relevance, and use Pearson correlation coefficient to measure redundancy between the genes.
- *ReliefF-mRMR*: It is two-stage selection algorithm by combining ReliefF and mRMR: In the first stage, ReliefF is applied to find a candidate gene set;

In the second stage, mRMR method is applied to directly and explicitly reduce redundancy for selecting a compact yet effective gene subset from the candidate set (Zhang et al., 2007).

- *F-statistic*: *F*-statistic is chosen to score the relevance between the genes and the target class. The *F*-statistic of gene *i* in *C* classes has the following form (Ding and Peng, 2003):

$$W_i = \frac{\sum_{c=1}^C n_c \cdot (\bar{g}_{ic} - \bar{g}_i)/(C-1)}{\sum_{c=1}^C \{(n_c - 1)[\sum_{j=1}^{n_c} (g_{jic} - \bar{g}_{ic})^2/n_c]/(n-C)\}} \quad (1)$$

where *C* is the number of classes,  $\bar{g}_i$  is the mean of gene *i* variables,  $n_c$  is the number of samples in class *c*,  $\bar{g}_{ic}$  is the mean of gene *i* in class *c*, and  $g_{jic}$  is sample *j* in gene *i* value in class *c*.

- *GNSR*: GSNR has been proposed and used in Zheng (2007) as a measure of the ratio between inter-group and intra-group variations. The GSNR value for gene *i* is given by:

$$W_i = \frac{\sum_{c=1}^C |\bar{g}_{jc} - \sum_{c=1}^C \bar{g}_{jc}/C|/C}{\sum_{i=1}^C \sum_{i=1}^{n_c} |g_{jic} - \bar{g}_{ic}|/n_c} \quad (2)$$

- *D-optimality and A-optimality*: They are model based approaches to estimate the information gain on the model (Yu et al., 2006; Zhu et al., 2008), instead of the data itself. A multivariate Gaussian generative model is used to fit the data and the criteria, *D*-optimality and *A*-optimality is used to select features (Fedorov, 1972). The criteria of *D*-optimality is to minimise the variance of the joint distribution of targets while that of *A*-optimality is to minimise average variance of all targets, or the trace of the covariance.
- *Other algorithms*: There are another eight selection measures (Su et al., 2003): *t*-statistic, twoing rules, information gain, gini index, max minority, sum minority, sum of variance, and one dimensional SVM.

## 1.2 Current gene selection packages

Currently, there are several existing gene selection software packages, such as rankgene (Su et al., 2003), and mRMR (<http://research.janelia.org/peng/proj/mrmr/index.htm>). Rankgene supports eight measures which are noted in other algorithms in the previous section for quantifying a gene's ability to distinguish between classes. And the existing version of mRMR only supports discrete variables such as MID and MIQ.

## 2 Implementation

Due to different developing environments of existing toolkits, it is difficult to compare different algorithms. We developed a software package named LIBGS in Matlab 7.0. LIBGS provides consistent input and output data formats for different algorithms, which make it more flexible to perform data analysis and algorithm comparison.

Furthermore, we add several new effective gene selection algorithms, such as ReliefF (Marko and Igor, 2003),  $F$ -statistic (Ding and Peng, 2003), GNSR (Zheng, 2007),  $A$ -optimality,  $D$ -optimality Fedorov (1972), and ReliefF-mRMR Zhang et al. (2007) to LIBGS.

### 2.1 Data structure and translation

LIBGS supports consistent data formats. Each gene dataset is formatted as a MATLAB data structure file(.mat), in which a class label vector corresponds to a gene array. For any algorithm, the input is a .mat file, and the output is an index vector for the selected genes. There are six gene data sets available in LIBGS, including ALL (<http://www.stjuderesearch.org/data/all1/>), GCM (Rifkin et al., 2001), HBC (<http://www.columbia.edu/~xy56/project.htm>), LYM (<http://genome-www.stanford.edu/lymphoma>), MLL (<http://research.dfci.harvard.edu/korsmeyer/ml1.htm>.) and NCI60 (<http://genome-www.stanford.edu/nci60/>). Furthermore, a utility is provided for converting the data from .csv file to .mat file. The command line is as follows.

```
csvtomat(Filename)
```

where Filename is the name of .csv file. In the .csv file, the first column is the class label, the rest are gene variables. For .mat file, its structure can be shown as Figure 1:

**Figure 1** Data structure description

```
v =
      X: [62x4026 double]
      y: [62x1 double]
      name: 'Finite data set'
      dim: 4026
      num_data: 62
```

We also provide the function to convert .mat file to .csv file as:

```
mattocsv(X, y, Filename)
```

where  $X, y$  are the matrix defined in .mat file and Filename is the .csv file as output file.

### 2.2 Implementation of gene selection algorithms

The command list to perform different gene selection algorithms is shown in Table 1, where  $X$  is a gene array,  $y$  is a class label vector, and Topn is the number of selected genes in current algorithm. In detail, for ReliefF function,  $n$  is the number of iterations,  $K$  is the number of neighbours to be selected, and typed is the data type; for Ftest-mRMR function,  $T1$  and  $T2$  are used together to choose different optimisation functions in mRMR; for  $A$ -optimality and  $D$ -optimality function,  $L$  is the regularisation term and for the Rankgene function, and  $T$  is the method index which can be referenced in rankgene.

**Table 1** MATLAB command list

<i>Algorithm description</i>	<i>Command line</i>
ReliefF	$W = \text{reliefF}(X,y,n,K,\text{typed},\text{Topn})$
<i>F</i> -statistic	$W = \text{Ftest}(X,y,\text{Topn})$
GNSR	$W = \text{Gsnr}(X,y,\text{Topn})$
Ftest-mRMR	$W = \text{mRMRC}(X,y,T1,T2,\text{Topn})$
ReliefF-mRMR	$W = \text{rm}(X,y,n,K,\text{Topn})$
<i>A</i> -optimality	$W = \text{fsAopt3}(X,y,L,\text{Topn})$
<i>D</i> -optimality	$W = \text{fsDopt3}(X,y,L,\text{Topn})$
Rankgene	$W = \text{rankgene}(X,y,T,\text{Topn})$

### 2.3 Assistant tools for classification

In addition, to compare the performance of the gene selection algorithms, we also include two popular classification tools in LIBGS, which are the existing MATLAB version for LIBSVM (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>) and a MATLAB Interface for WEKA (<http://www.cs.waikato.ac.nz/ml/weka/>). For LIBSVM, there is already a ready-to-run plug-in for MATLAB. And we implement the function for calling WEKA. The command line for calling WEKA is shown as follows.

```
mattocsv( $X, y$ , Filename)
Accuracy = wekaclassifier(Filename, Classifier)
```

where *Filename* is the name of the output .csv file,  $X$  is a gene array,  $y$  is a label vector, and *Classifier* is the parameter for classification method, such as Naive Bayes and J4.5 tree.

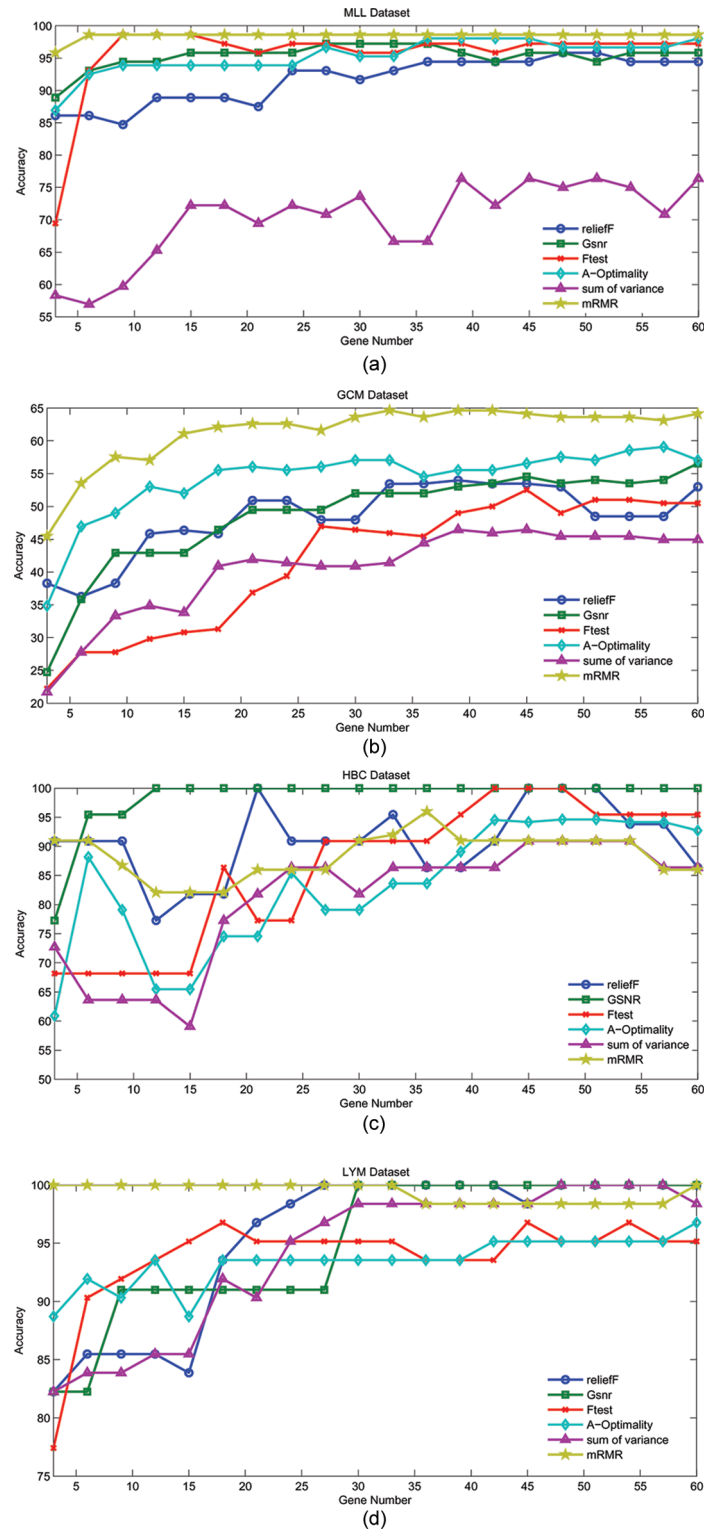
## 3 Results and discussion

Different methods can yield different lists of genes and lead to different results in classification. Figure 2 shows an example of performance comparison of three gene selection methods such as ReliefF, GSNR, *F*-statistic, *A*-optimality, sum of variance and mRMR-MID algorithms with the MLL, GCM, LYM, and HBC datasets using Naive Bayes Classifier. It is evident that with LIBGS, users can examine, compare the performance of different gene selection algorithms easily and efficiently.

## 4 Availability and requirements

LIBGS can be downloaded from <http://www.cs.fiu.edu/~yzhan004/genesel.html>. A detailed user guide and several examples are also available LIBGS is implemented and tested in Matlab 7.0. It can be integrated into the Toolbox by adding its path to MATLAB search path. In addition, the java version of WEKA should be install first and add weka.jar in your current directory.

**Figure 2** Comparison of ReliefF, Gsnr, Ftest, A-optimality, sum of variance and mRMR with: (a) MLL; (b) GCM; (c) HBC and (d) LYM datasets (see online version for colours)



## 5 Authors' contributions

Tao Li initialised the idea and supervised the project. Yi Zhang and Dingding Wang implemented the algorithms, developed the software, and performed experimental comparisons. Yi Zhang also built the website and wrote the user guide. Tao Li collected the datasets and revised the draft. All authors have read and approved the manuscript.

## 6 Conclusion

LIBGS is an integrated software package for performing, evaluating and comparing various gene selection algorithms. It can both output the information of selected genes and the comparison of classification accuracy using different gene selection approaches. Thus, LIBGS is a useful yet convenient tool in computational analysis of gene expression data.

## Acknowledgement

Tao Li is partially supported by an IBM Faculty Research Award, NSF CAREER Award IIS-0546280 and NIH/NIGMS S06 GM008205.

## References

- Ding, C. and Peng, H. (2003) 'Minimum redundancy feature selection from microarray gene expression data', *Proceedings of CSB'03*, pp.523–529.
- Fedorov, V.V. (1972) *Theory of Optimal Experiments*, Academic Press Inc.
- Li, T., Zhang, C. and Ogihara, M. (2004) 'A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression', *Bioinformatics*, Vol. 20, No. 15, pp.2429–2437.
- Marko, R.S. and Igor, K. (2003) 'Theoretical and empirical analysis of reliefF and rreliefF', *Machine Learning Journal*, Vol. 53, pp.23–69.
- Peng, H., Long, F. and Ding, C. (2005) 'Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy', *IEEE PAMI*, Vol. 27, pp.1226–1238.
- Su, Y., Murali, T.M., Pavlovic, V. and Kasif, S. (2003) 'Rankgene: identification of diagnostic genes based on expression data', *Bioinformatics*, Vol. 19, pp.1578–1579.
- Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., Poggio, T., Gerald, W., Loda, M., Lander, E.S., Golub, T.R., Ramaswamy, S. and Tamayo, P. (2001) 'Multiclass cancer diagnosis using tumor gene expression signatures', *Proceedings of National Academy of Sciences*, Vol. 98, pp.15149–15154.
- Ye, J., Li, T., Xiong, T. and Janardan, R. (2004) 'Using uncorrelated discriminant analysis for tissue classification with gene expression data', *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, Vol. 1, pp.181–190.
- Yu, K., Bi, J. and Tresp, V. (2006) 'Active learning via transductive experimental design', *ICML*, pp.1081–1088.
- Zhang, Y., Li, T. and Ding, C. (2007) 'A two-stage gene selection algorithm by combining reliefF and mRMR', *BIBE*, pp.164–171.

Author: Please supply location for highlighted references.

Zheng, G. (2007) *Statistical Analysis of Biomedical Data with Emphasis on Data Integration*, PhD Thesis, Florida International University.

Zhu, S., Wang, D. Yu, K. and Li, T. (2008) 'Feature selection for gene expression using model-based entropy', *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, Vol. 1, pp.1–2.

## **Websites**

<http://genome-www.stanford.edu/lymphoma>  
<http://genome-www.stanford.edu/nci60/>  
<http://research.dfci.harvard.edu/korsmeyer/ml1.htm>  
<http://research.janelia.org/peng/proj/mrmr/index.htm>  
<http://www.columbia.edu/~xy56/project.htm>  
<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>  
<http://www.cs.waikato.ac.nz/ml/weka/>  
<http://www.stjude.com/research/data/all1/>