AMB ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

A robust approach based on Weibull distribution for clustering gene expression data

Huakun Wang[1,2†], Zhenzhen Wang[1†], Xia Li[1*], Binsheng Gong[1], Lixin Feng[2] and Ying Zhou[2]

Abstract

Background: Clustering is a widely used technique for analysis of gene expression data. Most clustering methods group genes based on the distances, while few methods group genes according to the similarities of the distributions of the gene expression levels. Furthermore, as the biological annotation resources accumulated, an increasing number of genes have been annotated into functional categories. As a result, evaluating the performance of clustering methods in terms of the functional consistency of the resulting clusters is of great interest.

Results: In this paper, we proposed the WDCM (Weibull Distribution-based Clustering Method), a robust approach for clustering gene expression data, in which the gene expressions of individual genes are considered as the random variables following unique Weibull distributions. Our WDCM is based on the concept that the genes with similar expression profiles have similar distribution parameters, and thus the genes are clustered via the Weibull distribution parameters. We used the WDCM to cluster three cancer gene expression data sets from the lung cancer, B-cell follicular lymphoma and bladder carcinoma and obtained well-clustered results. We compared the performance of WDCM with k-means and Self Organizing Map (SOM) using functional annotation information given by the Gene Ontology (GO). The results showed that the functional annotation ratios of WDCM are higher than those of the other methods. We also utilized the external measure Adjusted Rand Index to validate the performance of the WDCM. The comparative results demonstrate that the WDCM provides the better clustering performance compared to k-means and SOM algorithms. The merit of the proposed WDCM is that it can be applied to cluster incomplete gene expression data without imputing the missing values. Moreover, the robustness of WDCM is also evaluated on the incomplete data sets.

Conclusions: The results demonstrate that our WDCM produces clusters with more consistent functional annotations than the other methods. The WDCM is also verified to be robust and is capable of clustering gene expression data containing a small quantity of missing values.

Background

The changes of the gene expression levels are very common in the human complex diseases, such as cancers [1-3]. The advent of microarray technologies have made it possible to measure simultaneously the expression levels of many thousands of genes over different time points and/or under different experimental conditions [4-6]. Numerous computational techniques have been developed to analyze these gene expression data. Among them, clustering is a primary approach to group the genes with similar expression patterns across different conditions, which enables the identification of differentially expressed gene sets in cancerous tissues [7-9]. Clustering is an unsupervised learning technique which assigns a set of objects (genes) into subsets (called *clusters*) so that the objects in the same clusters are similar according to some similarity metric [10,11]. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters.

Since clustering is proposed, an increasing number of clustering approaches have been developed and improved for the analyses of gene expression data. The

* Correspondence: lixia@hrbmu.edu.cn
† Contributed equally
[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, 150081, PR China
Full list of author information is available at the end of the article

common clustering methods include k-means [12,13], hierarchical clustering [8], and Self Organizing Map (SOM) [14,15], and so on. Each method has its own strengths and weaknesses. The k-means is an important clustering algorithm which partitions n objects into k clusters in which each object belongs to the cluster with the nearest mean. In k-means clustering, the number of clusters k is an input parameter, and an inappropriate choice of k may yield poor clustering results. The main advantages of this algorithm are its simplicity and computational speed which allows it to run on large datasets, however, it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments. Besides, it conducts poorly with overlapping clusters and is sensitive for noisy data. The hierarchical clustering aims to create a hierarchy of clusters which may be represented by a tree structure called a dendrogram. The root of the tree consists of a single cluster containing all objects, and the leaves correspond to individual objects. The hierarchical technique requires relatively smooth data and the clusters themselves need to be well defined. Like k-means method, noisy data strongly affect the resulting clusters. SOM is a type of artificial neural network that is trained using unsupervised learning to produce a two-dimensional, discretized representation of the input space of observations. It requires the geometry of nodes as input, and the nodes are mapped into two-dimensional space, initially at random, and then iteratively adjusted. SOM imposes the structure on data, with neighboring nodes tending to define related clusters. SOM has good computational properties and is suited to clustering of large data sets. One major drawback of this algorithm is the "boundary effect" of nodes on the edges of the network, which may lead to less effective clustering results. Besides, these clustering methods mentioned above require a complete data set as an input, and therefore those gene rows containing the missing values are either removed or imputed using an imputation method on the missing entries prior to clustering analysis. Removing the missing gene rows may result in omitting some important genes, such as the genes related to diseases, whereas the badly estimated missing values even changes the quality of data, which could influence the accuracy of clustering results.

In this article, we propose a Weibull distribution-based clustering method called WDCM. The assumption of this method is that the gene expression of each gene can be considered as a random variable following unique Weibull distribution [16], and that a group of genes tend to be clustered together if the Weibull distributions of gene expressions of these genes have similar distribution parameters. Here, we use the gene expression values of each gene to construct its corresponding Weibull distribution and then group these genes by clustering their corresponding distribution parameters.

The following sections of this paper are organized as 'Results', 'Discussion and conclusion' and 'Methods'. In section 'Results', we first introduced three cancer gene expression data sets we used, and then visually demonstrated the clustering results obtained using the WDCM for the three data sets. Second, to assess the performance of the WDCM, we compared the functional consistency of the gene clusters produced by the WDCM to those of the k-means and SOM methods for the same data sets. We also used the external measure Adjusted Rand Index to establish the performance of the WDCM, and the comparisons with the other algorithms were conducted simultaneously. Finally, we tested the robustness of the WDCM on clustering the incomplete data sets. In section 'Discussion and conclusion', we first summarized the main work of this study, discussed the strength and limitation of the WDCM. In the end we briefly mentioned the improvement of the WDCM and the future study. In section 'Methods', we introduced the WDCM together with the algorithm used for clustering the Weibull distribution parameters, the functional consistency assessment method of the clustering result, and the external validation index Adjusted Rand Index of the clustering performance. Moreover, Robustness test of the WDCM on clustering the incomplete data set was also presented in this section.

Methods

In this section, the WDCM is described as follows: Given a m × n gene expression matrix, let g_{ij} be the *jth* expression value of gene *i*, $i = 1, ...,m$, and $j = 1, ...,n$. We here treat one gene expression as a random variable, and construct the distribution of the gene expressions of gene *i*. We then choose a subset of genes whose distributions of the gene expressions belong to the common Weibull distribution [16]. Due to the consistent distribution function types, we consider that those genes with similar gene expression distribution parameters tend to share the similar expression patterns, and they are probably concerned with the same biological processes or functions together. We further cluster the genes in the selected subset by clustering their corresponding distribution parameters, as each gene corresponds to its unique distribution parameters. In the following we introduce the principle of the distribution function construction procedures.

Weibull distributions of gene expressions construction

First, we construct the empirical distribution of each gene expression [17], and then ascertain the precise distribution regarding the constructed empirical distribution using the Kolmogorov goodness of fit test [18-20].

The details as follows: assume that x_{i1}, x_{i2}, ..., x_{in} are the gene expressions of gene *gi*, i = 1, ...,m, and sort them in ascending as $x'_{i1} < x'_{i2} < \cdots x'_{in}$. For $\forall\ x \in (-\infty,+\infty)$, define the empirical distribution of g_i as

$$F_n^{(i)}(x) = \sum_{k=1}^{n} I(x'_{ik} \le x)/n \qquad (1)$$

Where $I(\cdot)$ is the indicator function.

We utilize the Weibull distribution type to fit $F_n^{(i)}(x)$, and then ascertain the distribution parameters which uniquely determine the distribution.

The probability density function of a Weibull distribution is defined as:

$$f(x;a,b) = \begin{cases} \dfrac{b}{a}\left(\dfrac{x}{a}\right)^{b-1} e^{-\left(\frac{x}{a}\right)^{b}} & , x \ge 0 \\ 0 & x < 0 \end{cases} \qquad (2)$$

where a >0 is the scale parameter and b >0 is the shape parameter of the distribution. The scale parameter a determines the range of the distribution. The shape parameter b is what gives the Weibull distribution its flexibility. By changing the value of the shape parameter, the Weibull distribution can fit a wide variety of data.

Let $\mathrm{F}^{(i)}(x)$ is a certain Weibull distribution with known parameters, and a Kolmogorov-Smirnov test is conducted to determine if the sample x_{i1},x_{i2}, ..., x_{in} comes from the Weibull distribution $\mathrm{F}^{(i)}(x)$. The null hypothesis is that the random sample of gene expressions of g_i comes from the Weibull distribution $\mathrm{F}^{(i)}(x)$. If the null hypothesis is true, the deviation of $\mathrm{F}^{(i)}(x)$ and $\mathrm{F}^{(i)}(x)$ is small. Construct the Kolmogorov-Smironov statistic

$$T_n^{(i)} = \sup_{x \in \Re} |F_n^{(i)}(x) - F^{(i)}(x)| \qquad (3)$$

under the null hypothesis, $\sqrt{n}T_n^{(i)}$ converges to the Kolmogorov distribution [18]. The null hypothesis is rejected at significance level α if $\sqrt{n}T_n^{(i)} > K_\alpha$, otherwise it is accepted, where K_α is the critical value of the Kolmogorov distribution. Given α = 0.05, we here select the appropriate parameters for $F^{(i)}(x)$ in order to the null hypothesis is accepted (p - $value$ > 0.05), that is, the random sample comes from the certain Weibull distribution $F^{(i)}(x)$, i = 1,2, ...,m. Following the above procedure, we can obtain the Weibull distributions of m gene expressions, denoted by $F^{(1)}(x)$,$F^{(2)}(x)$,...,$F(m)(x)$.

Weibull distribution parameters of gene expressions clustering

Let θ_i denotes the parameter of the Weibull distribution $F^{(i)}(x)$, j = 1, ...,m. Here θ_i consists of double-parameter pair (a_i,b_i), we then cluster the m parameters θ_1, θ_2,..., θ_m using a certain clustering algorithm based on the hub points. This algorithm presented by Robert Clason designates a single point as a hub for each cluster and then finds the distance from each remaining point to each hub, as well as assigns this point to the hub to which it is closer [21]. The merit of it is to automatically ascertain the clusters number on the basis of the distances between data points. A detailed description of the algorithm is provided in Additional file 1.

Functional consistency of clustering result

In order to evaluate the performance of the proposed WDCM, we also apply the K-means and Self Organizing Map (SOM) clustering algorithms to the same gene subsets as the WDCM and obtain the gene clusters, respectively. We compare the functional consistency of the gene clusters produced by WDCM to those produced by the other methods. For this purpose, we consider the biological annotations of the gene clusters in terms of Gene Ontology (GO). The Gene Ontology (GO) project provides three structured, controlled vocabularies that describe the gene products in terms of their associated biological processes (BP), cellular components (CC) and molecular functions (MF) [22]. The annotation ratios of each gene cluster in three GO terms were calculated using the web-accessible DAVID 2008 tool [23]. For each of clusters found by one of three clustering methods, under the BP ontology, we search the just GO term in which the most genes in this cluster are enriched, and define the BP annotation ratio for this cluster as the number of genes in both the assigned GO term and this cluster divided by the number of genes in this cluster. After calculating the BP annotation ratios for all clusters, we treat the mean value of all annotation ratios as the final BP annotation ratio. We also define the CC and MF annotation ratios by the same manner. A higher annotation ratio represents that the corresponding clustering result is better than the other ones, that is, gene are better clustered by function, indicating a more functionally consistent clustering result.

Adjusted Rand Index validation index

The Adjusted Rand Index (*ARI*) is a measure of agreement between two partitions of the same set of objects [24,25]. One partition is given by the clustering method and the other is defined by the external criteria. For a gene expression data set, suppose X is the partition based on some external criteria and C is the clustering result obtained by some clustering method. Let a,b,c,d respectively denote the number of gene pairs that are in the same cluster in both X and C, the number of gene pairs that are in the same cluster in X and in different clusters in C, the number of gene pairs that are in different clusters in X and in the same cluster in C and the number of gene pairs that are in different clusters in both X and C. The Adjusted Rand Index $ARI(X,C)$ is defined as follows:

$$ARI(X,C) = \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)} \tag{4}$$

The value of Adjusted Rand Index varies from 0 to 1 and higher value means that C is more similar to X.

Considering that the genes with similar expression patterns may be functionally related each other [26], we group the genes in the given data set according to functional similarity and define these gene clusters as X. The clustering results Cs are then given by the proposed WDCM, k-means and SOM. We compute and compare the values of Adjusted Rand Index between X and Cs to evaluate the performance of WDCM. To this end, we first use the Gene Functional Classification Tool of DAVID to group the genes into the highly functionally related gene clusters and then compute the values of *ARI*. The higher value indicates the corresponding clustering method performs better.

Robustness of the WDCM on clustering incomplete data set

The WDCM can be applied to cluster the incomplete gene expression data set without imputing the missing values. To test the robustness of this approach, we compared the overlapped degree between the gene clusters for incomplete data sets and the ones for complete data sets. A higher overlapped degree represents a robust clustering method. To this end, we first randomly remove 5-25% of the complete data set in order to create the incomplete gene expression data sets, and then we apply the WDCM to cluster these complete and incomplete data sets and obtain the clustering results, respectively. Here, a Cluster Overlap Ratio (COR) index is introduced for assessing the overlapped degrees at individual missing percentages.

Cluster Overlap Ratio index

Suppose n gene clusters $C_1,C_2,\ldots,C_n$ for the complete data set and m gene clusters $I_1,I_2, \ldots I_m$ for the incomplete one. The Cluster Overlap Ratio (COR) index is then defined as follows:

$$COR = \sum_{i=1}^{m} p_i x_i \tag{5}$$

where

$$p_i = \frac{|I_i|}{\sum_{k=1}^{m} |I_k|}, \tag{6}$$

$|\cdot|$ denotes the number of genes in the cluster, and thus p_i represents the proportion of genes in the gene cluster I_i. Here x_i denotes the maximum of overlapped gene numbers between I_i and each individual C_k (k = 1, ..., n) divided by $|I_i|$.

Results

Identification of six gene clusters for lung cancer data set

We applied the WDCM to cluster the lung cancer data set. It consists of expression levels of 675 genes across 156 tissues, which include 17 normal and 139 carcinomas lung tissues [27]. Using the Kolmogorov-Smirnov goodness of fit test (see Methods), we tested whether the expression sample of each gene comes from the Weibull distribution. The results showed that the distributions of gene expressions of 402 genes belong to the common Weibull distribution, whereas the others whose distributions of gene expressions fail to be in the Weibull distribution are removed. The *p-values* produced by Kolmogoriv-Smirnov goodness of fit test for the 402 genes were reported in Additional file 2. We then used the hub node based clustering algorithm (see Methods) to cluster the 402 Weibull distribution parameters which consist of the shape parameters and scale parameters, and obtained 6 distribution parameter clusters, that is, 6 gene clusters. The clustered parameters scatter plots have been shown in Figure 1A.

It is evident from Figure 1A that the distribution parameters of the genes of a cluster are close and compact to each other, which indicates the Weibull distribution parameters were clustered well. The expression profiles of the corresponding clustered genes plots have been shown in Figure 1B, from which it is also evident that the expression profiles of the genes within identical clusters are quite similar, whereas the profiles for the genes belonging to different clusters differ from each other.

Identification of four gene clusters for follicular lymphoma data set

We tested the WDCM on another follicular lymphoma data set consisting of expression levels of 798 genes in 19 B-cell follicular lymphoma specimens [28]. We utilized the Kolmogorov-Smirnov test to decide if the sample of individual gene on the follicular lymphoma data set comes from the Weibull distribution, and found 471 genes whose distributions of gene expressions belong to the common Weibull distribution. The *p-values* produced by Kolmogoriv-Smirnov goodness of fit test for the 471 genes were reported in Additional file 2. We then clustered the corresponding 471 distribution parameter pairs and determined 4 gene clusters. Figure 2 illustrates the clustered parameters scatter plots and the cluster profile plots of the clustering results.

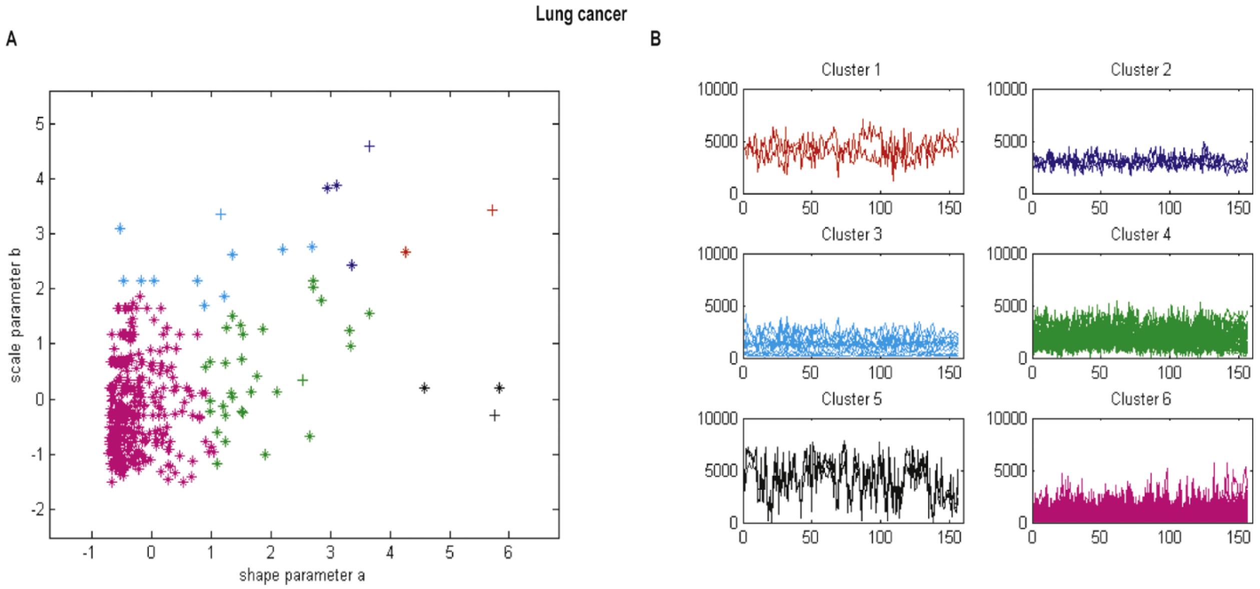

Figure 1 Lung cancer data set clustered using the WDCM. (A) Distribution parameters scatter plot. The horizontal axis corresponds to shape parameter a, and the vertical axis corresponds to scale parameter b. The parameter pairs in different clusters were drew with different colors. (B) Cluster profile plots.

From Figure 2A, the four parameters clusters are clearly distinguished from each other, meanwhile, the expression profiles of the genes within the same clusters are similar, whereas the ones of the genes across different clusters are distinct (see Figure 2B). The results indicate that the significantly distinct gene clusters were found using the WDCM on follicular lymphoma data set.

Identification of four gene clusters for bladder carcinoma data set

The bladder carcinoma data set contains 1203 genes measured over 40 different experimental conditions [29]. Using the Kolmogorov-Smirnov test, we found 1040 genes whose distributions of gene expressions belong to the common Weibull distribution. The *p-values* produced by Kolmogoriv-Smirnov goodness of

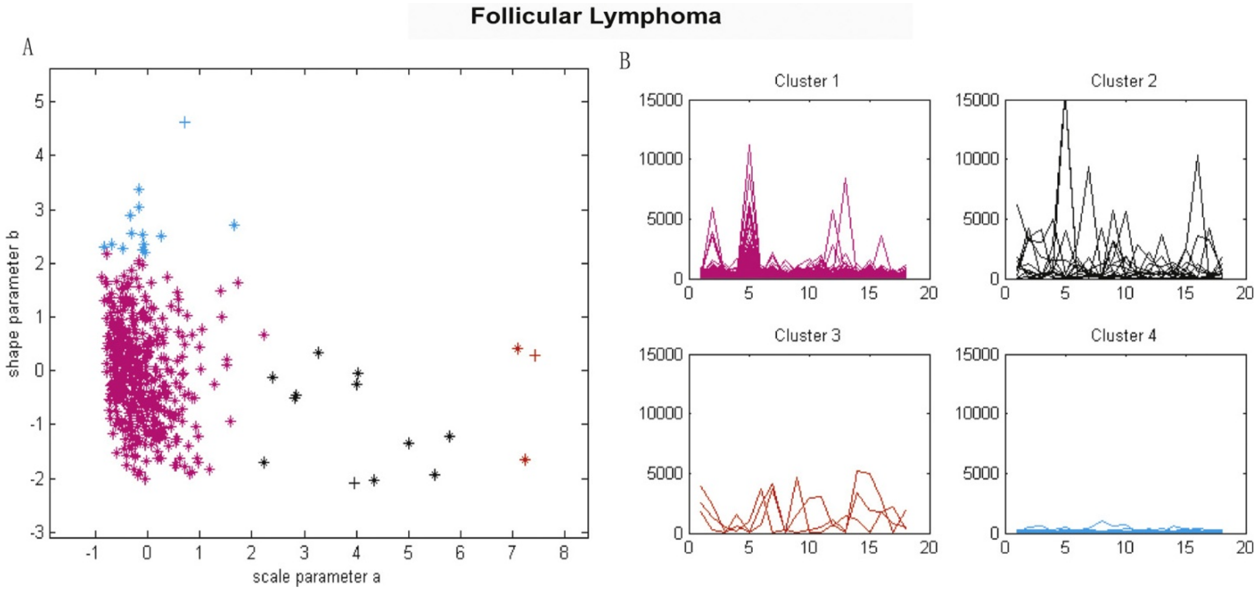

Figure 2 Follicular lymphoma data set clustered using the WDCM. (A) Distribution parameters scatter plot, (B) Cluster profile plots.

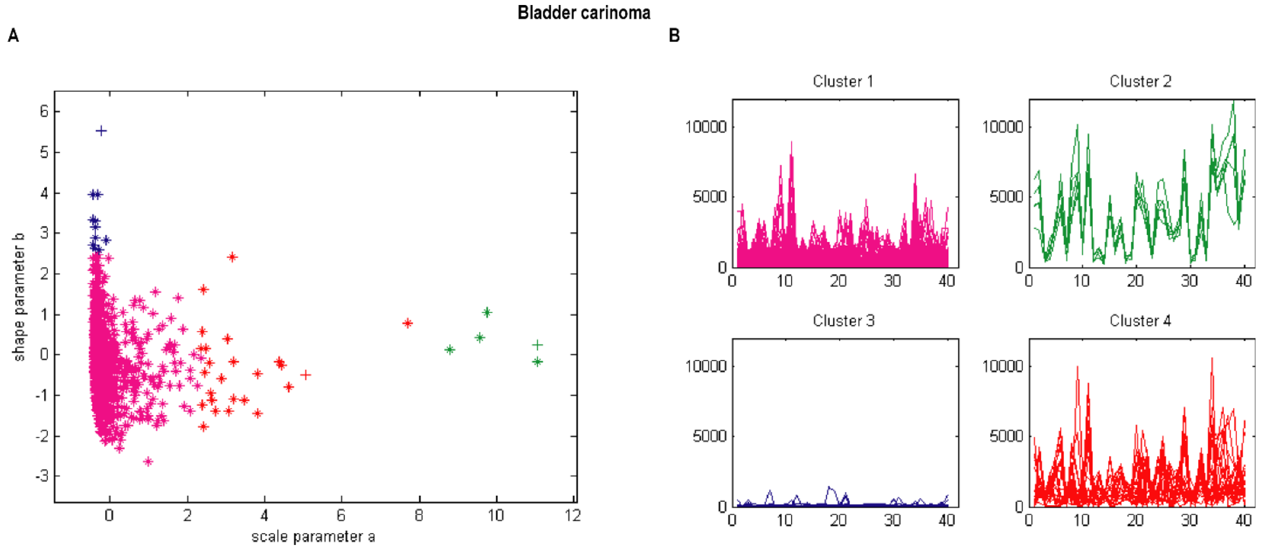

Figure 3 Bladder carcinoma data set clustered using the WDCM. (A) Distribution parameters scatter plot, (B) Cluster profile plots.

fit test for the 1040 genes were reported in Additional file 2. Again, the hub node based clustering algorithm was employed to cluster the corresponding 1040 distribution parameter pairs. The number of clusters determined was 4. Figure 3 shows the clustered parameters scatter plots and the cluster profile plots of the clustering results.

Comparison of clustering performance

To show the performance of the WDCM, we applied the K-means and Self Organizing Map (SOM) algorithms to the same gene subsets clustered by the WDCM and compared the functional consistency of the gene clusters produced by WDCM to those of the gene clusters produced by the other methods (see Methods). Simultaneously, the values of *ARI* for the WDCM, k-means and SOM algorithms on these three data sets were also contrasted (see Methods).

Among these three tested algorithms, the WDCM show the highest functional annotation ratios on both lung cancer and follicular lymphoma data sets. The detailed comparisons for the lung cancer data set are given in Figure 4A, from which we found that the three final functional annotation ratios of the WDCM clusters all exceed the ones of the other methods clusters. Especially, the BP and MF annotation ratios of the WDCM clusters (91.57% and 92.16%) are much higher than those of the SOM clusters (82.76% and 83.96%). On B-cell follicular lymphoma data test, although the CC and MF annotation ratios of gene clusters found by each of three methods are asymptotically equal (see Figure 4B), the BP annotation ratio of WDCM clusters (84.9%) is much higher than those of K-means clusters (71.6%) and SOM clusters (74.8%). On bladder carcinoma data set, from Figure 4C, although the BP annotation ratio of WDCM clusters (59.82%) is less than those of SOM clusters (64.30%), it is still beyond that of K-means clusters (55.87%). Note that the CC and MF annotation ratios of the WDCM clusters are consistently superior to those of the K-means and SOM clusters.

Table 1 shows the values of *ARI* for algorithms WDCM, k-means and SOM on these three data sets. Note that among the three methods, WDCM provides the consistently best *ARI* values. Specifically, the *ARI* value for the proposed WDCM (0.5365) is much better than those for k-means and SOM (0.2478 and 0.3681) on lung cancer data set. Although these three *ARI* values (0.3991, 0.3481 and 0.2647) are close on B-cell follicular lymphoma data set, the *ARI* value for WDCM is better than the other values. For bladder carcinoma data set also, the proposed WDCM outperforms the other algorithms in terms of *ARI*. The values are reported in Table 1.

The above comparative analyses on the functional annotation ratios of the three algorithms have demonstrated that the genes in each cluster obtained using the WDCM show not only the similar expression patterns, but also more consistent functional annotations, which means these genes are more inclined to be involved in the same biological functions together. Also, the Adjusted Rand Index comparative results indicate the superiority of the performance of the proposed WDCM compared to the other algorithms.

Test for robustness of the WDCM on clustering incomplete data set

To test the robustness with which the WDCM clusters the incomplete gene expression data, we applied the

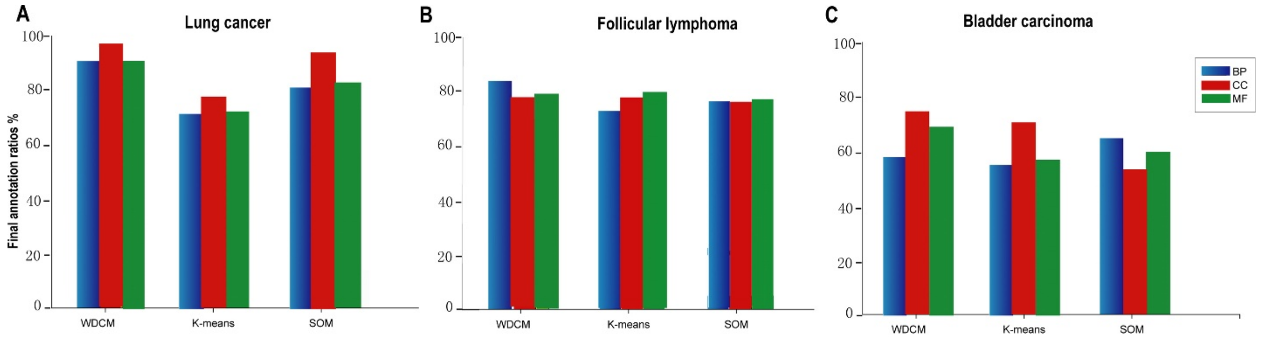

Figure 4 biological annotation ratios of clustering results. (A) Final annotation ratios of Lung cancer clusters found by three different methods in GO biological processes (BP), cellular components (CC) and molecular functions (MF). (B) Final annotation ratios of Follicular lymphoma clusters found by three different methods in GO biological processes (BP), cellular components (CC) and molecular functions (MF). (C) Final annotation ratios of Bladder carcinoma clusters.

WDCM to cluster the above three gene expression data sets containing missing values and compared the overlapped degree between the gene clusters for incomplete data sets and the ones for complete data sets. These three data sets were preprocessed by randomly removing 5-25% of the data in order to create the incomplete gene expression data sets, and the WDCM then was applied to these data sets. Table 2 lists the average Cluster Overlap Ratio (COR) values with respect to the percentages of missing values (0-25%) achieved by WDCM over 100 runs for the lung cancer, B-cell follicular lymphoma and bladder carcinoma data sets, respectively. The WDCM provided the higher COR values regarding the smaller percentages of missing values for all three data sets. The COR values exceeded 0.9 at 5% missing value. At 10%, the COR value was also beyond 0.9 for the follicular lymphoma and bladder carcinoma data sets (0.9078 and 0.9702), and approximated 0.9 for the lung cancer data set (0.8654). For the bladder carcinoma data set, we see that the COR values were varied from 0.9823 to 0.9335, passing 0.9 at all missing values.

The results of the cluster overlapped degree comparison tests indicate that the WDCM gave a high overlapped degree of the gene clusters compared with those of complete data set at low missing value, highlighting the robustness and potential of the WDCM. We think that the results might stem from the fact that the missing gene expression values of individual genes have little influence on constructing their corresponding Weibull distribution parameters at low missing values.

Table 1 ***ARI*** **values of WDCM, k-means and SOM algorithms for the lung cancer, B-cell follicular lymphoma and bladder carcinoma gene expression data sets**

Algorithm	Lung cancer	Follicular lymphoma	Bladder carcinoma
WDCM	**0.5365**	**0.3991**	**0.4105**
k-means	0.2478	0.3481	0.1623
SOM	0.3681	0.2647	0.0926

Discussion and conclusion

In this article, we propose a robust approach based on Weibull distribution (WDCM) for clustering gene expression data. It is based on the idea that a group of genes tend to be clustered together if the distributions of gene expressions of these genes belong to the common Weibull distribution and have the similar distribution parameters. Consequently, we cluster the genes by clustering the distribution parameters of their gene expressions. A hub nodes-based dynamic clustering algorithm is utilized in the distributions clustering process. The clusters number in a gene expression data set is automatically determined in this clustering algorithm. The performance of the proposed WDCM has been compared with those of K-means and SOM clustering algorithms by the biological annotation ratios to show its effectiveness on three cancer gene expression data sets. The results show that the WDCM is more capable of grouping the genes with similar expression patterns and strong functional consistency together. We also used the external measure Adjusted Rand Index to validate the performance of the WDCM. The comparative results demonstrate that the WDCM provides the better

Table 2 COR indices with respect to the specified percentages of missing values for the lung cancer, B-cell follicular lymphoma and bladder carcinoma data sets

Percentage of missing	Lung cancer	Follicular lymphoma	Bladder carcinoma
5%	0.9140	0.9495	0.9823
10%	0.8654	0.9078	0.9702
15%	0.8220	0.8738	0.9565
20%	0.7892	0.8418	0.9450
25%	0.7649	0.8120	0.9335

clustering performance compared to k-means and SOM algorithms. Moreover, the WDCM can be applied to cluster the incomplete gene expression data set without imputing the missing values. The results have demonstrated that there is high overlap between the gene clusters for the incomplete data set and those for the complete data set, which illustrates the robustness of the WDCM on clustering the incomplete data set at low percentage of missing values.

In general it is known that due to the complex nature of the gene expression data sets themselves and the experimental errors in detecting the gene expression data, it is difficult to discover an acknowledged best clustering approach. In clustering process, the WDCM disregards a few genes whose gene expression distributions fail to fit the Weibull distribution. In future study, we will consider replacing the single Weibull distribution with the mixture distribution in order to cluster the whole data set. Besides, we will also increase the robustness of this approach on clustering the incomplete gene expression data set containing the missing values of moderate percentage. For the gene clusters found by WDCM, we would like to investigate which gene clusters and genes are correlated with some cancer phenotype, and which biological processes or molecular functions these genes in the clusters are concerned with. Our study may be helpful to gain insights into the complex diseases.

Additional material

Additional file 1: A clustering algorithm based on "hub nodes". A clustering algorithm used to cluster the Weibull distribution parameters.

Additional file 2: P-values of tests for the three data sets. This file consists of three spreadsheets, each lists the gene numbers and p-values of Kolmogorov Smirnov test for one data set.

Acknowledgements
This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 30871394 and 61073136), the National High Tech Development Project of China, the 863 Program (Grant Nos. 2007AA02Z329), the National Basic Research Program of China, the 973 Program 9(Grant Nos. 2008CB517302) and the National Science Foundation of Heilongjiang Province (Grant Nos. JC200711, ZD200816-01), the Graduate Student Creation Science Foundation of Heilongjiang Province (Grants Nos. YJSCX2008-123HLJ) and the Scientific Research Foundation of Heilongjiang Provincial Education Department (Grants Nos. 11551362).

Author details
[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, 150081, PR China. [2]School of Mathematical sciences, Heilongjiang University, Harbin, 150080, PR China.

Authors' contributions
HKW and ZZW jointly proposed this approach and conducted the data experiments. XL gave the statistical idea of the method. BSG modified this paper. LXF partly wrote the program codes. Testing was done by YZ. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 24 December 2010 Accepted: 31 May 2011
Published: 31 May 2011

References

1. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO: Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* 2000, **24**:227-235.
2. Schlom J, Tsang KY, Kantor JA, Abrams SI, Zaremba S, Greiner J, Hodge JW: Cancer vaccine development. *Expert Opin Investig Drugs* 1998, **7**:1439-1452.
3. Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW: Gene expression profiles in normal and cancer cells. *Science* 1997, **276**:1268-1272.
4. Khademhosseini A: Chips to Hits: microarray and microfluidic technologies for high-throughput analysis and drug discovery. September 12-15, 2005, MA, USA. *Expert Rev Mol Diagn* 2005, **5**:843-846.
5. Khan J, Bittner ML, Chen Y, Meltzer PS, Trent JM: DNA microarray technology: the anticipated impact on the study of human disease. *Biochim Biophys Acta* 1999, **1423**:M17-28.
6. Watson A, Mazumder A, Stewart M, Balasubramanian S: Technology for microarray analysis of gene expression. *Curr Opin Biotechnol* 1998, **9**:609-614.
7. Ben-Dor A, Shamir R, Yakhini Z: Clustering gene expression patterns. *J Comput Biol* 1999, **6**:281-297.
8. Guess MJ, Wilson SB: Introduction to hierarchical clustering. *J Clin Neurophysiol* 2002, **19**:144-151.
9. Rahnenfuhrer J: Clustering algorithms and other exploratory methods for microarray data analysis. *Methods Inf Med* 2005, **44**:444-448.
10. Boutros PC, Okey AB: Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data. *Brief Bioinform* 2005, **6**:331-343.
11. Sierra A, Corbacho F: Reclassification as supervised clustering. *Neural Comput* 2000, **12**:2537-2546.
12. MacQueen JB: Some Methods for classification and Analysis of Multivariate Observations. *the 5th Berkeley Symposium on Mathematical Statistics and Probability* University of California Press; 1967, 281-297.
13. Gourevitch B, Le Bouquin-Jeannes R: K-means clustering method for auditory evoked potentials selection. *Med Biol Eng Comput* 2003, **41**:397-402.
14. Cottrell M, Ibbou S, Letremy P: SOM-based algorithms for qualitative variables. *Neural Netw* 2004, **17**:1149-1167.
15. Lee BH, Scholz M: Application of the self-organizing map (SOM) to assess the heavy metal removal performance in experimental constructed wetlands. *Water Res* 2006, **40**:3367-3374.
16. Weibull W: A statistical distribution function of wide applicability. *J Appl Mech-Trans ASME* 1951, **18**:293-297.
17. Turnbull BW: The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society Series B* 1976, **38**:290-295.
18. Frank J, Massey J: The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association* 1951, **46**:68-78.
19. Huang S, Yeo AA, Li SD: Modification of Kolmogorov-Smirnov test for DNA content data analysis through distribution alignment. *Assay Drug Dev Technol* 2007, **5**:663-671.
20. Ong LD, LeClare PC: The Kolmogorov-Smirnov test for the log-normality of sample cumulative frequency distributions. *Health Phys* 1968, **14**:376.
21. Clason R: Finding Clusters: An application of the Distance Concept. *The Mathematics Teacher* 1990.
22. Blake JA, Harris MA: The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis. *Curr Protoc Bioinformatics* 2008, **7**, **Unit 7 2**.
23. Huang da W, Sherman BT, Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009, **4**:44-57.
24. Yeung KY, Haynor DR, Ruzzo WL: Validating clustering for gene expression data. *Bioinformatics* 2001, **17**:309-318.

25. R Giancarlo DS, Utro F: *Statistical Indexes for Computational and Data Driven Class Discovery in Microarray Data. In Biological Data Mining* Chapman and Hall; 2009.
26. Mosca E, Bertoli G, Piscitelli E, Vilardo L, Reinbold RA, Zucchi I, Milanesi L: **Identification of functionally related genes using data mining and data integration: a breast cancer case study.** *BMC Bioinformatics* 2009, **10(Suppl 12)**:S8.
27. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.** *Proc Natl Acad Sci USA* 2001, **98**:13790-13795.
28. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR: **Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.** *Nat Med* 2002, **8**:68-74.
29. Dyrskjot L, Thykjaer T, Kruhoffer M, Jensen JL, Marcussen N, Hamilton-Dutoit S, Wolf H, Orntoft TF: **Identifying distinct classes of bladder carcinoma using microarrays.** *Nat Genet* 2003, **33**:90-96.

doi:10.1186/1748-7188-6-14
Cite this article as: Wang *et al.*: A robust approach based on Weibull distribution for clustering gene expression data. *Algorithms for Molecular Biology* 2011 **6**:14.