BOHR Journal of Biocomputing and Nano Technology
2025, Vol. 3, No. 1, pp. 9–17
DOI: 10.54646/bjbnt.2025.16
www.bohrpub.com

RESEARCH

Principal directon divising partitioning initialisation of K-means clustering for discriminating among leukemias while identifying the most salient genes involved

Diego Liberati[*†]

National Research Council of Italy, Politecnico di Milano University, Milan, Italy

[*]**Correspondence:**
Diego Liberati,
diego.liberati@cnr.it

[†]**ORCID:**
Diego Liberati
0000-0002-4294-6705

Received: 12 September 2025; **Accepted:** 14 October 2025; **Published:** 01 November 2025

This paper attempts to cluster leukemia patients described by gene expression data and to discover the most discriminating genes that are responsible for the clustering. A combined approach of Principal Direction Divisive Partitioning (PDDP) and bisecting K-means algorithms is applied to the clustering of the investigated leukemia dataset. Both unsupervised and supervised methods are considered in order to get optimal results. The combination of PDDP and bisecting K-means successfully clusters leukemia patients and efficiently discovers salient genes able to discriminate the clusters. The combined approach works well on the automatic clustering of leukemia patients depending merely on the gene expression information, and it has great potential for solving similar problems, like classifying pancreatic tumors. The salient identified genes may thus enhance relevant information for discriminating among leukemias. A previous paper by us, cited in the references and in the paper, based on the same technique, was able to outperform a seminal paper on Science on their same data. In this paper, the bisection is iterated on more complex data in order to identify a tree of leukemias discriminated through their salient involved genes.

Keywords: leukemias, genes, clustering, K-means, principal component analysis

Introduction

The rapid development of the deoxyribonucleic acid (DNA) microarray technology is making it more and more convenient to obtain various gene expression datasets with abundant information that can be very helpful for many meaningful biomedical applications such as prediction, prevention, diagnosis, and treatment of diseases; development of new drugs, patient-tailored therapy, and precision and personalized medicine. However, these datasets are usually very large and unbalanced, with the number of genes (thousands upon thousands) being much greater than the number of patients (generally from tens to hundreds). Consequently, how to analyze effectively this kind of large dataset with few samples and numerous attributes—for example, how to classify according to their gene expression profile the patients suffering from certain diseases, or how to determine from thousands of genes the most discriminating ones that are responsible for the corresponding disease—should beviewed as an important issue.

Literature review

In the recent decades there have been many exciting research results in the area of DNA microarray data mining on the basis of gene expression data analysis. For instance, to cite a few pioneer results, depending solely on gene

expression monitoring to microarray datasets, Golub et al. (1) classified sample patients of acute leukemia as two subtypes, Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML), and predicted the subtypes of new leukemia cases according to the expression values of the most decisive genes that were discovered during the classification of sample cases; Armstrong et al. (2) discovered a new subtype of acute leukemia, Mixed Lineage Leukemia (MLL), claimed as distinct enough to be separated from ALL or AML; From a hierarchical point of view, De Cecco et al. (3) classified patients of advanced ovarian cancer stage 2 and extracted significant genes that characterized each level in the hierarchies; On the basis of gene expression profile analysis van't Veer et al. (4) predicted the clinical outcome (relapse/non-relapse) of breast cancer and Pomeroy et al. (5) predicted the outcome (survivor/failure) of embryonal tumor of central nervous system; Alon et al. (6) clustered correlated gene families about colon tissues and separated cancerous from non cancerous tissues; Singh et al. (7) performed the tumor versus normal classification of prostate cancer and predicted the clinical outcome of prostatectomy; Yeoh et al. (8) classified the sub types and predicted the outcome of pediatric ALL; Gordon et al. (9) separated malignant pleural mesothelioma (MPM), which is not a lung cancer, from adenocarcinoma (ADCA) of the lung; Alizadeh et al. (10) identified two distinct types of diffuse large B-cell lymphoma (DLBCL), the germinal centre B-like DLBCL and the activated B-like DLBCL.

The technologies applied in the analysis of gene expression data are various. In Golub et al. (1), a method of neighborhood analysis is used to select out the most informative genes that are related to the classification of patients, a class predictor is designed by using the sum of the weighted votes from these genes to determine the winning class, and a cross-validation method is adopted to test the accuracy of the predictor. To classify the leukemia patients, a technology of self-organizing maps is applied to obtain two classes. In van't Veer et al. (4), an unsupervised method is used to cluster both genes and tumors, and a supervised alternative is adopted to identify the outcome of the tumors and extract the most significant genes that are related to the outcome. In Pomeroy et al. (5), Principal Component Analysis (PCA) is applied to determine different types of tumors and the related genes. In Alon et al. (6), a deterministic-annealing algorithm is used to organize both genes and sample tissues into binary trees so that they can be clustered hierarchically. In Gordon et al. (9), gene expression ratios are calculated, and thresholds are selected to distinguish between cancer and non-cancer tissues.

Structure and aim of the present paper

In this paper, an approach based on the collaboration of three algorithms, PCA, Principal Direction Divisive Partitioning (PDDP), and bisect K-means, is applied to cluster the sample patients from a public leukemia dataset (2) consisting of 72 leukemia samples—24 ALL, 20 MLL, and 28 AML, each sample being represented by 12,582 gene expression values. In the meantime, the few significant genes more determinant to the clustering results are identified.

Methodology

Description of the dataset

The dataset analyzed in this paper is the combination of two leukemia datasets processed in Armstrong et al. (2), where 57 samples (20 ALL, 17 MLL, and 20 AML) are used for training and 15 (4 ALL, 3 MLL, and 8 AML) for testing the clustering of leukemia patients. Each patient is determined by a sequence of 12,582 real numbers, each measuring the relative expression of the corresponding gene. The data set can then be viewed as 72 points in a 12,582-dimensional Euclidean space. A simple measure of the genomic difference between two patients can be obtained by resorting to the Euclidean distance of two points. In order to ease the algebraic manipulations of data, the dataset can also be represented as a real 2-D matrix S of size $72 \times 2{,}582$; the entry s_{ij} of S measures the expression of the jth gene of the ith patient.

Pre-processing of the dataset

The leukemia dataset is a very large matrix with more than 10,000 genes as its columns, while a great portion of them, with small changes of values between different patients, provides much less information related to the patient clustering than the residual small portion, in which large differences of values can be found between different patients or patient types. In this dataset, it can be observed that a very large portion of genes has relatively small standard deviation values, although the values vary from 0 to 15,000. For example, at least 10,000 standard deviation values are less than 1,200. Therefore, prior to the patient clustering, it is possible to apply a filter to remove those genes of little importance (11). In order to analyze such a huge dataset without any filters, a higher amount of time and storage would be needed, as well as a larger amount of computational resources. The removing of less important genes can help decrease the complexity of analysis and the requirement of computational resources without much affecting result precision. Furthermore, the removing of those genes may also reduce the interference caused by noise.

By taking all these factors into account, a pre-processing of the dataset is applied first to remove those genes with small standard deviation values. A threshold of 400 is used to filter out the genes with standard deviation values less than it. The

dataset after this preprocessing almost halves, becoming a 72 × 6,611 matrix with the removal of 5,971 gene columns. The reason for using 400 as the threshold is that it keeps a large portion of the data so that the important information will reasonably not be ignored, at the same time removing another large portion—almost a half—of the data to speed up the clustering procedures. In the following sections, unless otherwise specified, all the analysis is based on the 72 × 6,611 dataset after the pre-processing with threshold th = 400.

Description of algorithms

The clustering analysis of the leukemia dataset is based on three steps. First, with the principal component analysis, all the genes in the dataset are sorted according to their significance to the patient clustering. Then, the dataset is clustered using a modified bisect K-means algorithm, which is essentially the combination of PDDP, used to initialize the following, and K-means. Finally, the minimum set of genes minimizing clustering errors is identified. This gene set consists of a few necessary and sufficient genes in the sense of the clustering approach applied in this paper, but the so-identified genes are also keen to provide useful information for the differential diagnosis and even better understanding of the corresponding subtypes of leukemia.

PCA

It is well known that the PCA method (12–14) works better on measuring the contribution of attributes of samples to the clustering when the dataset can be linearly partitioned. The extraction of principal components is briefly recalled as follows for the sake of completeness: Given a p × N dataset S where p and N are, respectively, the numbers of samples and attributes. If dataset S is a centralized matrix where each column (i.e., attribute) of S has zero mean value, then the first principal component of S should be the eigenvector corresponding to the largest eigenvalue of the covariance matrix of S, namely S_{TS}; the second principal component of S should be the eigenvector corresponding to the second largest eigenvalue of S_{TS}, and so on. A simple proof is given out in (12).

The principal components can be obtained from the singular value decomposition (SVD) (14) of S as the product of three special matrices. When a principal component, generally the one corresponding to the largest singular value, is selected out, the degree of contribution of the attributes to the clustering of samples can be quantified by comparing the absolute values of the elements in the principal component vector. The positions of the largest absolute values point out the most discriminating attributes for clustering the sample.

When the dataset matrix S is not centralized, with the mean values of some attributes being non-zeros, the SVD should be performed on the centralized form of S so as to equally weight the contribution from each attribute.

PDDP

The PDDP algorithm is proposed by Boley (15). It has the following steps: (1) For the matrix S (in general S is not centralized) in Section "The unsupervised clustering of dataset S", first calculate the mean value vector $w = [w_1, w_2, \ldots, w_N]$ for all the samples. The mean value vector is the centroid of the samples. (2) Calculate matrix S_0, the centralized form of S. (3) Select an appropriate principal component $v = [v_1, v_2, \ldots, v_N]T$ for S_0, where vector v is determined either manually or automatically by the method described in Section "The selection of principal components". (4) Write matrix S as $[S_1, S_2, \ldots, S_p]T$. If $(S_i\text{-}w)v \leq 0$, then S_iS_L, otherwise S_iS_R, where $1 \leq i \leq p$.

The rationale of PDDP has a geometrical interpretation. The dataset is first transformed to an N-dimensional coordinate system originating at the dataset centroid and having all the N component vectors (even if not principal) as coordinates. Suppose a principal component is selected to do PDDP; then the data points are separated as two clusters by an (N-1)-dimensional hyperplane passing through the origin and normal to such principal component vector. Generally speaking, some distance-based methods—such as the minimum distance and the average distance between two different clusters—can be used to measure the difference between them.

It should be pointed out that PDDP can be applied repeatedly to any cluster to get two subclusters; therefore, any number of clusters can be obtained by iteratively using such an algorithm. Savaresi et al. (16) have proposed a method to tell which one of two given clusters is more suitable to be further split, while Kruengkrai et al. (17) have suggested how to determine whether a cluster could again be split, thus helping to terminate iterations.

The selection of principal components

A possible problem of principal component selection. The selection of an appropriate principal component is the precondition of the success of PDDP clustering. In general, the first principal component is appropriate because it represents the primary direction of the dataset, and the direction itself is the very foundation of the PDDP algorithm. However, the first principal component may not always be a good choice. In this case the primary direction of the data points is still indicated by the first principal component, but obviously another principal component splits the dataset much better; therefore, this principal component, even though it is not the first one but just the second one, should be selected as the input of the PDDP algorithm.

The automatic selection of principal components. The selection of a principal component is easy for supervised PDDP clustering, because we can simply find out from a set of given candidates, for example, the first three principal components, the best one yielding the result closest to the reference. However, when an unsupervised PDDP clustering

algorithm is applied, the selection of an appropriate principal component should be done on an automatic basis other than manually. In Savaresi et al. (16), a method that is originally designed for selecting what clusters are to split is deemed to be also helpful for selecting principal components, just after slight modification.

Combining PDDP with bisecting K-means

The K-means algorithm performs well when the distance information between data points is important to the clustering. However, K-means has an intrinsic disadvantage. The clustering result depends greatly on the selection of initial "center points." Pang-ning Tan et al. (18) show different results by applying K-means on the same dataset with different choices of initial "center points." PDDP has its own weakness, too. Since the partition of PDDP is only on the basis of the projection from the data points to a selected principal direction, the distance information between these data points is ignored.

In spite of the fact that in many cases neither PDDP nor K-means alone is good enough for deriving desirable clustering results, according to the theory of Savaresi and Boley (19), Savaresi et al. (20), and Savaresi and Boley (21), the combination of PDDP and bisect K-means keeps the merits of both algorithms and usually performs better than either single one does. PDDP, although weak at taking advantage of distance information, can provide bisect K-means good initial center points that are close to true ones; therefore, the accuracy of bisecting K-means with clustering can be improved. The difference between the combined method and the traditional bisect K-means lies in the selection of the initial center points, c1 and c2. With the combined method, the two center points of bisect K-means are not selected randomly but according to the clustering result of PDDP; that is to say, c1 and c2 should be the sample mean values of the PDDP clusters S_L and S_R, respectively. The combination of PDDP and bisecting K-means makes the selection of c1 and c2 more reasonable by reducing the risk caused by a random selection.

The extraction of significant attributes

As already mentioned, the extraction of significant features strongly related to clustering is also a key issue, besides the clustering itself. To achieve this, one should first know the degree of significance of each attribute. Fortunately, principal component analysis itself can also provide quantitative information to measure the significance.

Supervised and unsupervised clustering

With a supervised clustering approach, some a priori knowledge, such as a predefined reference result and the number of clusters, can be used to guide the process of clustering. However, such a priori knowledge is not always available before clustering; they may be known only when the clustering is successfully completed. In this case, an unsupervised alternative can be considered when applicable. The PDDP + bisect K-means algorithm is capable of dividing data points into two clusters in either a supervised or unsupervised way.

Experimental case: data and results

This section is focused on some experimental results about the clustering of the leukemia gene expression dataset mentioned previously. The original dataset S consists of 72 samples (24 ALL, 20 MLL, and 28 AML patients distributed in a training dataset of 57 samples and a testing dataset of 15 samples), and each sample is represented by 12,582 gene expression values. The samples are numbered as follows: #1 - #20 (ALL in training), #21 - #37 (MLL in training), #38 - #57 (AML in training), #58 - #61 (ALL in testing), #62 - #64 (MLL in testing), and #65 - #72 (AML in testing). Dataset S is stored as a 72 × 12,583 matrix, because there is an extra column, column 12,583, which represents the clustering result presented in Armstrong et al. (2). In this column, classes ALL, MLL, and AML are represented as 0, 1, and 2, respectively. This column serves as the reference result of all the following experiments. In other words, the experiment results are compared with the reference, and any different clustering cases are reported as "errors" and analyzed later. As already said, before any experiments, a threshold th = 400 is applied to remove those genes with standard deviation values less than 400, since they are with little possibility to be significant attributes. To verify the effectiveness of the threshold, every experiment is then repeated with th = 0, i.e., all the genes included. The exactly same results and much less execution time show that the threshold applied is reasonable and effective, at least in this experimental case. All the experiments are based on the MATLAB implementation of the algorithms described in Section "Methodology".

The unsupervised clustering of dataset S

With threshold th = 400, the input dataset S becomes a 72 × 6,611 matrix. With the first principal component and all 6,611 genes, a clustering result in two initial clusters, S_L and S_R. It should be mentioned that this initial clustering successfully separates ALL and AML with only an exception at sample #3, if we claim that all ALL samples belong to S_L and all AML belong to SR. This implies that we correctly identify 23 out of 24 ALL and all 28 out of 28 AML samples, like in Garatti et al. (11).

The minimum gene set that produces the above result consists of only two genes:

#28 (in the original 12,582-attribute dataset), whose name is AFFX-UMGAPDH/M33197_5_at, and #12,430, with the name 256_s_at.

The significance coefficient gives information about these two genes. The significance coefficients are obtained by taking the absolute values of the corresponding elements in the first principal component, the average coefficient is the mean of the absolute values of all the 6,611 coefficients, and the normalized coefficients, which are used as the contribution indicator of the genes to the clustering, are the quotients of the significance coefficients to the average coefficient.

By considering the 72 expression values of these two genes (**Figure 1**), one can visually separate S_L (with relatively low expression values) from SR (with relatively high expression values) to a certain extent, although a few exceptional cases exist. The rationale of the extraction of these two genes is thus illustrated in such a manner.

It is natural that the initial clustering does not give out any useful information about the MLL samples, because the PDDP-based approach only produces two clusters after a single application. For this reason, further clustering is needed to hopefully reveal the aspect of the MLL part.

The unsupervised clustering of sub dataset S_L

According to the result of the initial clustering and bisection, 37 samples are classified as S_L; among them, 23 are actually ALL samples and 14 are MLL. Clustering of subclass S_L is continued in order to see if the PDDP-based approach can successfully identify these ALL samples from the non-ALL ones (i.e., according to the reference, the MLL ones at the first bipartition clustered with ALL, thus closer to such ones than to AML). With the first principal component, 5,962 genes (threshold th = 400), and again just two (other) significant genes, 33412_at and 769_s_at, a result is obtained exactly reproducing the reference.

We thus can claim that S_{LL} and S_{LR} are actually ALL and a part of MLL, respectively. **Figure 2** plot the 37 expression values of these two genes.

The unsupervised clustering of subdataset S_R

Since the initial clustering of dataset S is not sufficient for identifying the MLL samples, a similar clustering of the subclass S_R is then performed to see whether those MLL samples can be separated successfully. According to the result of the initial clustering, 35 samples are classified as S_R. Among them are 28 AML, 6 MLL, and one misclassified ALL. With the first principal component and 6,191 genes (threshold th = 400), the minimum gene set consists of 219 genes, too many to be reported in this paper.

The clustering seems thus unsuccessful at first glance, with many AML samples and all the MLL samples clustered together into S_{RL}. However, interestingly enough, no MLL sample is clustered into S_{RR}.

The supervised clustering of subdataset S_{RL}

Because all 6 MLL samples are classified as S_{RL} in Section "The unsupervised clustering of subdataset S_R", it may be interesting to continue clustering the subcluster S_{RL}. With the first principal component and 5,877 genes (threshold th = 400), an unsupervised result with two errors is obtained.

The minimum gene set for this result consists of 103 genes, still too many to be reported in this paper. However, when the clustering is performed under the supervision of the reference result, a better outcome is obtained with only one error at patient #3. The minimum gene set for this result consists of only nine genes:

319_g_at 0.1106
AFFX-HSAC07/X00351_5_at
AFFX-HSAC07/X00351_M_at
33412_at
33516_at
AFFX-HUMGAPDH/M33197_5_at
35083_at
36122_at
39318_at

Discussion

Discussion about the clustering results

According to the clustering results in Section "Experimental case: data and results", the leukemia dataset S can be clustered as the following hierarchy:

In **Figure 3**, if we name cluster S_{LL} as ALL, clusters S_{LR} and S_{RLL} together as MLL, and clusters S_{RLR} and S_{RR} together as AML, then there is only one error occurring in the whole set of patients with such supervised aggregation. Almost all the 24 ALL patients are identified in cluster S_{LL}, except patient #3, who is eventually misclassified into cluster S_{RLL}; this is the only error that occurs. It may be due to imprecision of the algorithm, but also to orginal misclassification of the data as in Golub et al. (1), as discovered by Garatti et al. (11), or it may just be a borderline subject difficult to classify being close to another class, at least in the reduced used subspace (11). 18 AML patients are identified in cluster SRR and the other 10 in cluster S_{RLR}; these two clusters include all the AML patients without any misclassification.

Interestingly enough, except for ALL, both MLL and AML patients are divided into two subclusters. This implies that there might exist other subtypes for MLL and AML. In fact, on the very same subset, Golub et al. (1) labeled only two

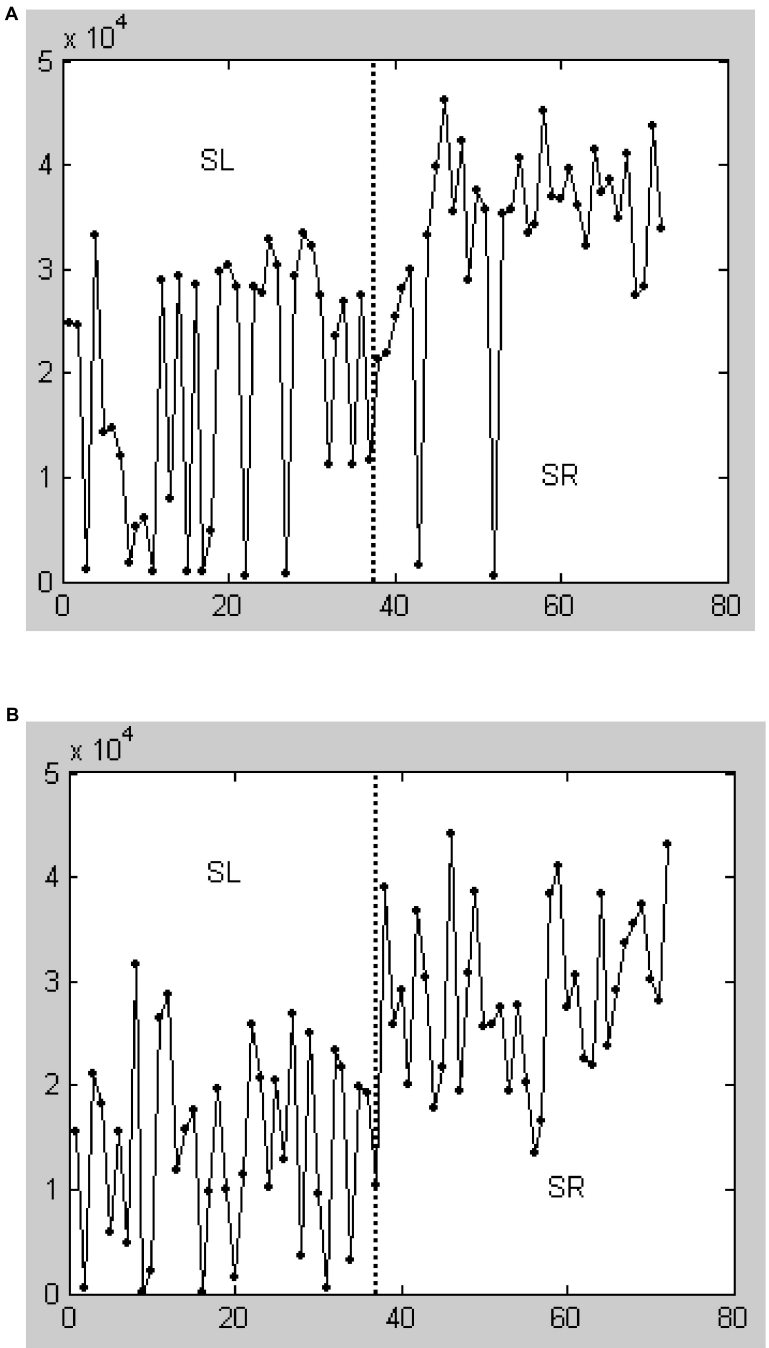

FIGURE 1 | (A) Expression values of gene #28. **(B)** Expression values of gene #12,430.

subtypes of leukemia (ALL and AML), while Armstrong et al. (2) detailed proposing the three subtypes (ALL, MLL, and AML) analyzed in this paper. In Grassi et al. (22), we were able, for instance, to confirm a suspected pattern in a further rare leukemia.

Discussion about the significant genes

First, by reviewing the gene extraction results in Section "Experimental case: data and results", we see that different levels of expression values of just the two genes #28 (AFFX-HUMGAPDH/M33197_5_at) and #12,430 (256_s_at) are already enough to well separate ALL and AML patients. Second, in the initial clustering of the dataset, most MLL data are shown closer to ALL than AML, implying that MLL and ALL share similarity to a great extent: in fact, they were classified together in the same class by Golub et al. (1). The difference between ALL and MLL is then very well revealed by just two more genes, #7,754 (33412_at) and #11,924 (769_s_at).

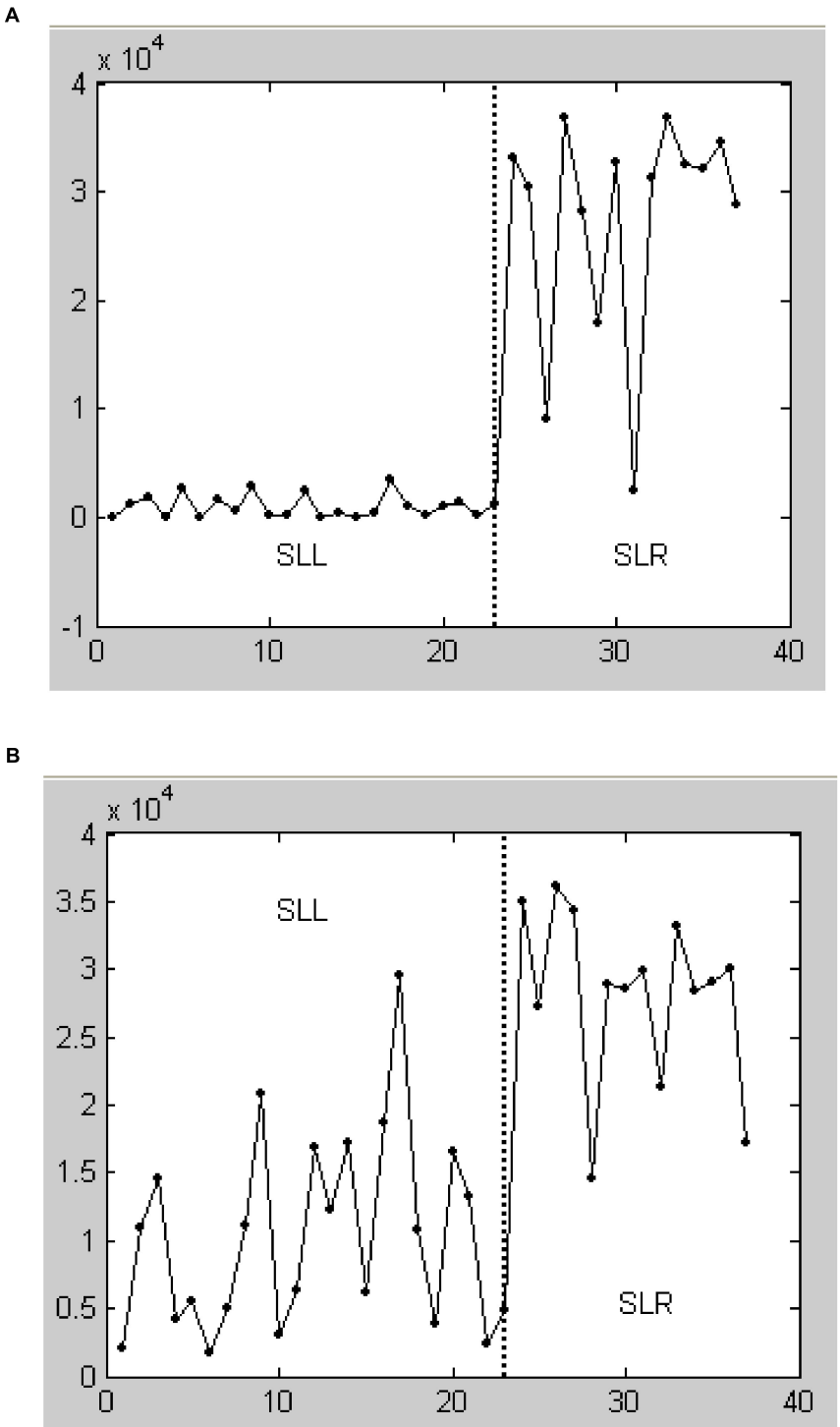

FIGURE 2 | (A) Expression values of gene #7,754. **(B)** Expression values of gene #11,924.

On the other hand, a small portion of MLL data are shown closer to AML, showing that some MLL and AML cases may have common characteristics. The size of the minimum set of genes that separate MLL from AML is very large, implying that genetically diagnosing AML-like MLL patients may be more difficult than that of ALL-like MLL patients. Finally, the contribution of genes to the corresponding clustering results is quantified so that the significance of them can be compared quantitatively. For example, genes #28 (normalized significance coefficient (NSC) = 15.2466) and #12,430 (NSC = 13.4795) are almost equally significant to the discrimination between ALL and AML, while gene #7,754 (NSC = 21.2917) appears to be more significant than #11,924 (NSC = 15.0472) to the discrimination between MLL and ALL, and so on.

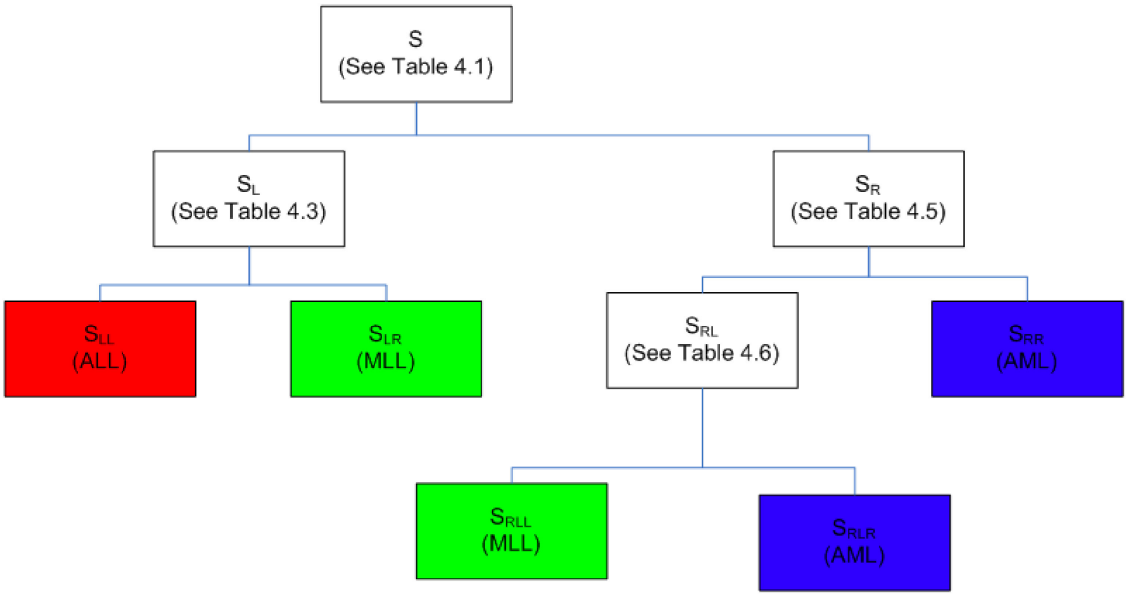

FIGURE 3 | The Hierarchy of the Leukemia Dataset, where the significant final results are in the colored blocks, the white ones being just their aggregates referring to omitted tables for the sake of reading.

Conclusion

In summary, via the combined approach of PDDP and bisect K-means, 72 leukemia patients are successfully clustered as ALL, MLL, and AML, respectively. Among all the 12,582 genes, the most discriminating ones that are responsible for the clustering are efficiently discovered. Furthermore, both the clustering of patients and the discovering of significant genes are performed automatically to a great extent and depend merely on the gene expression data, which can be obtained conveniently by using the popular DNA microarray technology.

In conclusion, the combination of PDDP and bisect K-means does confirm itself, as in Garatti et al. (11) and Grassi et al. (22), to be an efficient approach for the clustering of the leukemia patient dataset described in this paper, and hopefully also efficient for other similar datasets. Moreover, the significant genes discovered among tens of thousands of genes may provide very important information for the diagnosis of leukemia. The same approach reveals itself to be useful to other tumor classifications, like pancreatic ones (in preparation), even if not in all cases: it does not work as well, for instance, in discriminating breast cancer, but it is anyway able to provide useful insights, as it will be shown in a companion paper already submitted. This is understandable by considering that the proposed approach works in a quasi-linear partitioning that is not in general appropriate to any data set. When working, like in this paper, it offers a powerful, simlple approach to gain immediate knowledge about the few genes mainly involved in classification, thus possibly offering hints in understanding pathophysiology, as well as in suggesting and monitoring therapy that is beyond the scope of this very paper but inspiring for further research.

Funding

Research was not funded.

Acknowledgments

To colleagues, postdocs and students involved in the project at various titles and stages.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science.* (1999) 286:531–7.
2. Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet*. (2002) 30:41–7.
3. De Cecco L, Marchionni L, Gariboldi M, Reid JF, Lagonigro MS, Caramuta S, et al. Gene expression profiling of advanced ovarian cancer: characterization of a molecular signature involving fibroblast growth factor 2. *Oncogene*. (2004) 23(49):8171–83.
4. van't Veer LJ, Dai HY, van de Vijver MJ, He YDD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature.* (2002) 415:530–6.
5. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature.* (2002) 415:436–42.

6. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A*. (1999) 96:6745–50.
7. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, et al. Gene expression correlates of clinical prostate cancer behaviour. *Cancer Cell*. (2002) 1:203–9.
8. Yeoh E-J, Ross ME, Shurtleff SA, Kent Williams W, Patel D, Mahfouz R, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*. (2002) 1:133–43.
9. Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Ramaswamy S, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using Gege expression ratios in lung cancer and mesothelioma. *Cancer Res*. (2002) 62:4963–7.
10. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. (2000) 403: 503–11.
11. Garatti S, Bittanti S, Liberati D, Maffezzoli A. An unsupervised clustering approach for leukaemia classification based on DNA micro-arrays data. *Intell Data Anal*. (2007) 11(2):175–88.
12. Hand D, Mannila H, Smyth P. *Principles of Data-Mining*. Cambridge, Massachusetts, USA: The MIT Press (2001).
13. O'Connel MJ. Search program for significant variables. *Comput Phys Commun*. (1974) 8:49–55.
14. Wall ME, Rechtsteiner A, Rocha LM. Singular value decomposition and principal component analysis. In: Berrar DP, Dubitzky W, Granzow M editors. *A Practical Approach to Microarray Data Analysis*. Norwell, MA: Kluwer (2003). p. 91–109.
15. Boley DL. Principal direction divisive partitioning. *Data Min Knowl Discov*. (1998) 2(4):325–44.
16. Savaresi S, Boley D, Bittanti S, Gazzaniga G. Choosing the cluster to split in bisecting divisive clustering algorithms. *Second SIAM International Conference on Data Mining (SDM'2002)*. (2002).
17. Kruengkrai C, Sornlertlamvanich V, Isahara H. Refining a divisive partitioning algorithm for unsupervised clustering. *The 3rd International Conference on Hybrid Intelligent Systems (HIS'03)*. (2003).
18. Tan P, Steinbach M, Kumar V. *Introduction to Data Mining*. Boston: Addison Wesley Publishing Company (2005).
19. Savaresi SM, Boley DL. On the performance of bisecting K-means and PDDP. *1st SIAM Conference on Data Mining, Chicago, IL, USA, paper n. 5*. (2001). p. 1–14.
20. Savaresi SM, Boley DL, Bittanti S, Gazzaniga G. Cluster selection in divisive clustering algorithms. *2nd SIAM International Conference on Data Mining, Arlington, VI, USA*. (2002). p. 299–314.
21. Savaresi SM, Boley DLA. Comparative analysis on the bisecting K-means and the PDDP clustering algorithms. *Int J Intell Data Anal*. (2004) 8(4):345–62.
22. Grassi S, Palumbo S, Mariotti V, Liberati D, Guerrini F, Ciabatti E, et al. The WNT pathway is relevant for the BCR-ABL1-independent resistance in chronic myeloid leukemia. *Front Oncol*. (2019) 9:532.