A Gene Selection Approach based on Clustering for Classification Tasks in Colon Cancer

José A. Castellanos-Garzón[a,b] and Juan Ramos[c]

[a]Department of Computer Engineering, CISUC-ECOS, Faculty of Science and Technology, University of Coimbra, Pólo II - Pinhal de Marrocos, 3030-290 Coimbra.

[b]IEEE Member.

[c]Instituto de Investigación Biomédica de Salamanca (IBSAL), Complejo Asistencial Universitario de Salamanca, Hospital Virgen de la Vega, Planta 10, Paseo de San Vicente 58-182, 37007 Salamanca. jantonio@acm.com

KEYWORD

Gene selection; DNA-microarray; Clustering; Filter method; Colon cancer.

ABSTRACT

Gene selection (GS) is an important research area in the analysis of DNA-microarray data, since it involves gene discovery meaningful for a particular target annotation or able to discriminate expression profiles of samples coming from different populations. In this context, a wide number of filter methods have been proposed in the literature to identify subsets of relevant genes in accordance with prefixed targets. Despite the fact that there is a wide number of proposals, the complexity imposed by this problem (GS) remains a challenge. Hence, this paper proposes a novel approach for gene selection by using cluster techniques and filter methods on the found groupings to achieve informative gene subsets. As a result of applying our methodology to Colon cancer data, we have identified the best informative gene subset between several one subsets. According to the above, the reached results have proven the reliability of the approach given in this paper.

1. Introduction

Colorectal cancer (CRC) is the third most common type of malignancies worldwide and the second cause of cancer death among adults (Kim et al., 2015; Markowitz and Bertagnolli, 2009). Since cancer is a process related to aging, this relation is particularly strong in CRC (Haraldsdottir et al., 2014) where the incidence increases with increasing age in a remarkable way. CRC emerges from mutations in genes affecting a series of major carcinogenic pathways. 60% of all CRCs emerge from adenomas via the suppressor pathway, initiated by an APC gene mutation (Jass, 2007). About 5% of all cases of this disease is caused by a hereditary syndrome (Balaguer, 2014). Thus, the relevance of this disease is beyond doubt and demands further research for its better understanding (Perea et al., 2011). At its origin, CRC is a benign adenomatous polyp and before turning into an invasive cancer, it gradually progresses to an adenoma (Markowitz and Bertagnolli, 2009).

On the other hand, the high CRC incidence along with the variety of precursor mutations have changed the concept in the scientific community and it no longer refers to a single disease (Perea et al., 2011). Therefore, the molecular complexity taking place in CRC constitutes a major problem for clinic research. For this reason CRC requires, as in other cancers, further molecular characterization in order to help researchers who are concerned with the discovery of biomarkers. Thus, advances leading to the understanding of CRC molecular processes result essential for suitable knowledge management by part of both medical and researcher personnel (Perea et al., 2011).

Meanwhile, the study of certain cellular processes involved in diseases as cancer by means of microarray technologies has allowed us to monitor the expression levels for tens of thousands of genes in parallel (Geoffrey et al., 2004; Jiang et al., 2004; Berrar et al., 2003; Bourne and Wissig, 2003). From a simple experimental device given by DNA-microarray technology, researchers can monitor interactions of thousands of gene transcriptions in an organism. This technology is particularly useful in the evaluation of gene expression patters during important biological processes and across collections of related samples. So DNA-microarrays are able to observe at the same moment and in response to the same stimulus, the gene expression levels of many genes under different samples. However, criteria from molecular biology state that *only a small subset of genes participate in any cellular process of interest and that a cellular process takes place only in a subset of the samples* (Jiang et al., 2004). This belief has led to the emergence of the research area of *gene selection*.

Gene selection is leading to gene discovery relevant for a particular target annotation. Hence, those discovered genes play an important role in the analysis of gene expression data since they are able to differentiate samples from different populations (Natarajan and Ravi, 2014; Shraddha and Anuradha, 2014; Tyagi and Mishra, 2013; Lazar et al., 2012; Wang et al., 2005; Inza et al., 2004). Such genes are called *informative genes* or *differentially expressed genes*. Research in this field is primarily targeted at identifying biomarkers and classifying tissue samples based on *machine learning*. The application of clustering techniques to this area has proven to be helpful to identify informative genes (Lazar et al., 2012). In that sense, this paper proposes clustering-based approach for gene selection from DNA-microarray data. The strategy followed by this approach is to apply clustering and visualization techniques to partition the target dataset into smaller subsets and filter out the most significant genes from each cluster for classification tasks. Thus, we assess the found subsets of informative genes by training and measuring the accuracy of a classifier based on a *support vector machine* (SVM) (Martens et al., 2007; Han and Kamber, 2006; Vapnik, 1995). The informative gene subset achieving the best results on the classifier is selected as the most representative for the analyzed dataset.

To reach the goals above, the remainder of this paper has been divided into the following sections. Section 2 describes the methodology developed by our approach to acquire several sets of informative genes. Section 3 provides specific characteristics of the used Colon dataset as results and discussion after applying the gene selection approach. Section 4 gives the conclusions of the research and at the end of this paper, the used references have been outlined.

2. Material and methods

This section describes the approach proposed for gene selection from DNA-microarray data. The methodology developed in this approach has been summarized in Figure 1, which outlines four main parts being explained in the following subsections.

2.1 Hierarchical clustering analysis (HCA-I)

This stage is responsable for applying three hierarchical clustering methods to the target dataset. Thus, HCA-I returns three dendrograms as a result to the next stages. The selected clustering methods have been Agnes, Diana and Eisen (Kaufman and Rousseeuw, 2005; Eisen et al., 1998). These methods have widely been used in the literature in the analysis of DNA-microarray data and they have proven perform-well in this data domain (Castellanos-Garzón and Díaz, 2013; Castellanos-Garzón et al., 2013; Jiang et al., 2004; Berrar et al., 2003). So that, the Agnes algorithm builds a hierarchy of clusterings where each data is first a small cluster by itself. Clusters are merged until only one large cluster remains containing all data. At each stage the two nearest clusters are combined to form one larger cluster. The Diana algorithm performs the task in the reverse order, starting from a large cluster containing all data. Clusters are divided until each cluster contains only a single data. At each

stage, the cluster with the largest diameter is split. The Eisen algorithm carries out an agglomerative hierarchical clustering in which each cluster is represented by the mean vector from data in the cluster.

2.2 Clustering validation (CV-II)

This stage is responsable for selecting a clustering from each dendrogram given as input. To do this, CV-II focuses on cluster validity measure *silhouette width* (Kaufman and Rousseeuw, 2005) to evaluate each clustering in dendrograms. Silhouette width is a composite index, reflecting compactness and separation of clusters and can also be applied on different metrics.

Then, measure silhouette width is applied to clusterings of each dendrogram in order to select the best clusterings. Additionally, the selected clusterings as its corresponding dendrograms are explored by a cluster visualization tool (3D-VisualCluster, (Castellanos-Garzón et al., 2013)) to validate the made choice. Thus, as a result of the visual analysis carried out on such dendrograms, a new clustering can be selected for each dendrogram. Finally, three clusterings are returned as a result to the next stage. Note that the aim of applying clustering techniques is to achieve a first partition of data into similarity groups in such a way that a gene filtering method can be applied to a each group regardless.

2.3 Filtering process (FP-III)

This stage is what establishes the actual filtering process of genes. So for each clustering the most significant genes from its clusters are selected based on two filtering strategies. These strategies are applied to clusters and a new clustering is achieved by informative genes selected in each cluster. Then, each clustering is analyzed to identify its small clusters (clusters with a size less than or equal to eight genes) and big clusters (the other clusters). Since genes in small clusters tend to be correlated, only a single gene is selected with the highest variance. For the remaining big clusters in which their genes are less correlated, filter method Signal-to-Noise (S2N) (Lazar et al., 2012; Berrar et al., 2003) is applied. S2N computes the statistic that determines the correlation of each gene with respect to both tissue sample classes given in the dataset. Thus, the most positive values are more correlated with the positive class whereas the most negative values are more correlated with the negative class. Hence, a determined number of genes is selected for each class and finally, three informative gene clusterings (or sets) are passed to the next stage.

2.4 Classification process (CP-IV)

This is the last stage of our approach which is responsable for evaluating the quality of the informative gene sets given by the previous stage. In this regard, we use a linear support vector machine (SVM) as a classifier in order to test and evaluate different gene subsets. To do this, each subset of informative genes (in this case, three subsets of informative genes will be obtained, one for each clustering method) is used separately to train the classifier. Because there are not available test data to compute the accuracy of the classifier, we have adopted methodology *stratified 10-fold cross-validation* for each gene subset. The gene subset reaching the best accuracy will be selected as the most representative subset of informative genes for the given dataset. Therefore, the classifier built by such a set can be used to classify unknown sample patterns into one of the classes given in the analyzed dataset.

3. Results and discussion

This section outlines the dataset used in the experiment as the results of each stage of the applied methodology whereas their discussion is provided. So as said in section Introduction, we have focused on a DNA-microarray

dataset of colorectal cancer (CRC-dataset) available at http://genomics-pubs.princeton.edu/oncology/affydata/index.html. Gene expression in 40 tumor and 22 normal colon tissue samples from 40 patients has been processed on an Affymetrix oligonucleotide array complementary for more than 6500 human genes (Alon et al., 1999). Finally, a gene expression matrix with 2000 gene probes × 62 tissue samples has been achieved and normalized to a mean 0 and variance 1.

Then, as a result of applying stages I and II (HCA-I and CV-II) of the proposed approach to CRC-dataset, Table 1 has been achieved. This table displays the results of applying measure silhouette width to dendrograms given by the three used clustering methods. Columns Mean-SilhoW and Best-SilhoW list the mean (with standard error) and best silhouette width respectively reached by the corresponding dendrogram. Columns Best-Level and Chosen-Level refer to the number of clusters corresponding with column Best-SilhoW and the one of the level finally selected using cluster visualization techniques. Note that the level given by the silhouette width measure (column Best-Level) is just to know where to start exploring the dendrogram in search for the most suitable level (clustering) through the visualization tool. Thus, the clustering finally selected from each dendrogram in stage-II has been given in column Chosen-Level.

Table 1: Comparative table evaluating the silhouette width for dendrograms of methods Agnes, Diana and Eisen applied to CRC-dataset. The number of clusters for the best level and the one for level finally selected in the dendrogram have also been shown.

Method	Mean-SilhoW	Best-SilhoW	Best-Level	Chosen-Level
Agnes	0.519 ± 0.009	0.912	2-clusters	44-clusters
Diana	0.456 ± 0.011	0.843	2-clusters	75-clusters
Eisen	0.489 ± 0.012	0.912	2-clusters	62-clusters

Reinforcing the decision taken in column Chosen-Level in Table 1, Figures 2 and 3 show different visualizations of the dendrograms and clusterings given in this table. Figure 2 displays the CRC-dataset point distribution through 3D-scatterplots, representing each clustering from Table 1. clusters have been represented by gene-points with the same color. On the other hand, note that the CRC-dataset point distribution is much more compact on the right side and much more sparse on the left side of the point cloud of the dataset. This suggests that the tendency of a clustering algorithm applied to these data is to find a few big and median clusters (compact part) whereas most found clusters are unitary or very small (sparse part) as shown in this figure. The above can also be seen in Figure 3, where the clustering selected on each heatmap presents the characteristics above. Also keep in mind that very small clusters in the heatmaps of Figure 3 have not been stressed.

Going to the following and final stages, FP-III and CP-IV respectively, we have that the filtered process has been applied to the three achieved clusterings and three subsets of informative genes have been filtered out to be passed to CP-IV. Then, the three gene subsets have been used to train a SVM-classifier in order to evaluate its accuracy according to each subset. Thereby, Table 2 shows the number of genes in each subset and its result in the classifier by a test of stratified 10-fold cross-validation. As shown in the table, the three informative gene sets have reached the same accuracy, which means that according to the used dataset, such sets have been able to capture the genes essential to differentiate samples into different classes of CRC-dataset. Then, since the accuracy is the same for all gene sets given in Table 2, we have selected the informative gene set represented by the Agnes method as the most meaningful to classify CRC-dataset because it is the smallest, 76 genes.

Table 2: Comparative table showing the number of genes for each subset of informative genes along with its accuracy reached by the SVM-classifier of CRC-dataset.

Method	Gene-Subset	SVM-Accuracy
Agnes	76	88.710
Diana	146	88.710
Eisen	90	88.710

4. Conclusions

This paper has presented a clustering-based approach for gene selection from DNA-microarray data. The methodology followed by this approach has been tested on a public dataset of colorectal cancer (CRC-dataset). According to this, the aim of this research has then been to find the best subset of informative genes from several clustering results given by using methods developing different strategies on how to cluster the data. In this regard, we have used two strategies of gene filtering (see Subsection 2.3) fitted the kind of processed cluster (small or big clusters). Finally, the results reached from each clustering method have been evaluated on a linear classifier of a support vector machine and the best subset of informative genes has been selected as the most significant to represent each tissue sample of CRC-dataset. Thus, the methodology developed in this paper has proven to perform-well on DNA-microarray data by reaching results very promising on CRC-dataset. So the approach can be extended to other datasets of the current data domain.

Conflict of interest

The authors declare that they have no competing interests.

Contributors

JACG and JR, designed and wrote the proposed approach. JACG implemented the framework while JR provided biological background. JACG provided the comments and the discussion. All authors read and approved the final manuscript.

Appendix-A: Gene names of the best informative gene set found from CRC-dataset.

Informative gene list: Hsa.3004 H55933, Hsa.13491 R39465, Hsa.1447 T55131, Hsa.20836 R02593, Hsa.37254 R85482, Hsa.3087 T65938, Hsa.2555 X63432, Hsa.467 H20709, Hsa.2357 T52342, Hsa.750 T72863, Hsa.4689 T95018, Hsa.6080 J02763, Hsa.1737 T72175, Hsa.31 T57780, Hsa.140 M87789, Hsa.5398 T58861, Hsa.44472 H80240, Hsa.474 L28809, Hsa.909 M11799, Hsa.1977 T51496, Hsa.2597 T49423, Hsa.3002 R22197, UMGAP control, Hsa.1013 T61661, Hsa.8068 T57619, Hsa.5444 T48804, Hsa.957 M26697, Hsa.6039 H22688, Hsa.45604 H88360, Hsa.624 M57710, Hsa.538 T56940, Hsa.285 T62972, HSAC07 control, Hsa.2794 T48904, Hsa.1130 Z24727, Hsa.8147 M63391, Hsa.43279 H64489, Hsa.8374 T59162, Hsa.1119 T59954, Hsa.1836 T51574, Hsa.1139 T88723, Hsa.316 M94132, Hsa.831 M22382, Hsa.2800 X55715, Hsa.3016 T47377, Hsa.2361 T51534, Hsa.1013 T61661, Hsa.120 D14662, Hsa.3152

D31885, Hsa.3306 X12671, Hsa.773 H40095, Hsa.7395 R10066, Hsa.31630 R64115, Hsa.11673 H23544, Hsa.2451 U22055, Hsa.2705 X56597, Hsa.10664 T83368, Hsa.601 J05032, Hsa.462 U09564, Hsa.7 H72234, Hsa.2196 M58050, Hsa.3331 T86473, Hsa.3263 U26312, Hsa.2645 X54942, Hsa.549 R36977, Hsa.36689 Z50753, Hsa.37937 R87126, Hsa.1832 J02854, Hsa.692 M76378, i control, Hsa.8831 T49941, Hsa.891 M19045, Hsa.1039 M27190.

5. References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences, USA*, 96:6745–6750.

Balaguer, F., 2014. Cáncer colorrectal familiar y hereditario. *Gastroenterología y Hepatología*, 37:77–84.

Berrar, D. P., Dubitzky, W., and Granzow, M., 2003. *A Practical Approach to Microarray Data Analysis*. Kluwer Academic Publishers, New York, Boston, Dordrecht, London, Moscow.

Bourne, P. E. and Wissig, H., 2003. *Structural Bioinformatics*. Wiley-Liss, Inc., Hoboken, New Jersey.

Castellanos-Garzón, J. A. and Díaz, F., 2013. An Evolutionary Computational Model Applied to Cluster Analysis of DNA Microarray Data. *Expert Systems with Applications, Elsevier*, 40:2575–2591.

Castellanos-Garzón, J. A., García, C. A., Novais, P., and Díaz, F., 2013. A Visual Analytics Framework for Cluster Analysis of DNA Microarray Data. *Expert Systems with Applications, Elsevier*, 40:758–774.

Eisen, M., Spellman, T., Brown, P., and Botstein, D., 1998. Cluster Analysis and Display of Genome-wide Expression Patterns. *Proceedings of the National Academy of Sciences, USA*, 95:14863–14868.

Geoffrey, J. M., Do, K. A., and Ambroise, C., 2004. *Analyzing Microarray Gene Expression Data*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Han, J. and Kamber, M., 2006. *Data Mining: Concepts and Techniques*. Elsevier Inc.

Haraldsdottir, S., Einarsdottir, H., Smaradottir, A., Gunnlaugsson, A., and Halfdanarson, T., 2014. Colorectal cancer-review. *Laeknabladid*, 2(100):75–82.

Inza, I., Larrañaga, P., Blanco, R., and Cerrolaza, A., 2004. Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial Intelligence in Medicine, Elsevier*, 31:91–103.

Jass, J., 2007. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology*, 50:: 223–230.

Jiang, D., Tang, C., and Zhang, A., 2004. Cluster Analysis for Gene Expression Data: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1370–1386.

Kaufman, L. and Rousseeuw, P. J., 2005. *Finding Groups in Data. An Introduction to Clustering Analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Kim, S.-E., Paik, H., Yoon, H., Lee, J. E., Kim, N., and Sung, M.-K., 2015. Sex- and gender-specific disparities in colorectal cancer risk. *World Journal of Gastroenterology: WJG*, 17(21):5167–5175.

Lazar, C., Taminau, J., Meganck, D., S.and Steenhoff, Coletta, A., Molter, V., C.and deSchaetzen, Duque, H., R.and Bersini, and Nowé, A., 2012. A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis. *IEEE/ACM Transactions On Computational Biology And Bioinformatics*, 9(4):1106–1118.

Markowitz, S. and Bertagnolli, M., 2009. Molecular basis of colorectal cancer. *New England Journal of Medicine*, 25(361):2449–2460.

Martens, D., Baesens, B., Van, T. G., and Vanthienen, J., 2007. Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, 183:1466–1476.

Natarajan, A. and Ravi, T., 2014. A Survey on Gene feature selection using microarray data for cancer classification. *International Journal of Computer Science & Communication (IJCSC)*, 5(1):126–129.

Perea, J., Lomas, M., and Hidalgo, M., 2011. Molecular basis of colorrectal cancer: towards an individualized management. *Revista Española de Enfermedades Digestivas*, 1(103):29–35.

Shraddha, S. and Anuradha, S., N.and Swapnil, 2014. Feature Selection Techniques and Microarray Data: A Survey. *International Journal of Emerging Technology and Advanced Engineering*, 4(1):179–183.

Tyagi, V. and Mishra, A., 2013. A Survey on Different Feature Selection Methods for Microarray Data Analysis. *International Journal of Computer Applications*, 67(16):36–40.

Vapnik, V., 1995. *The nature of statistical learning theory*. Springer, New York.

Wang, Y., Tetko, I., Hall, M., Frank, E., Facius, A., Mayer, K., and Mewes, H., 2005. Gene selection from microarray data for cancer classification - a machine learning approach. *Computational Biology and Chemistry, Elsevier*, 29:37–46.

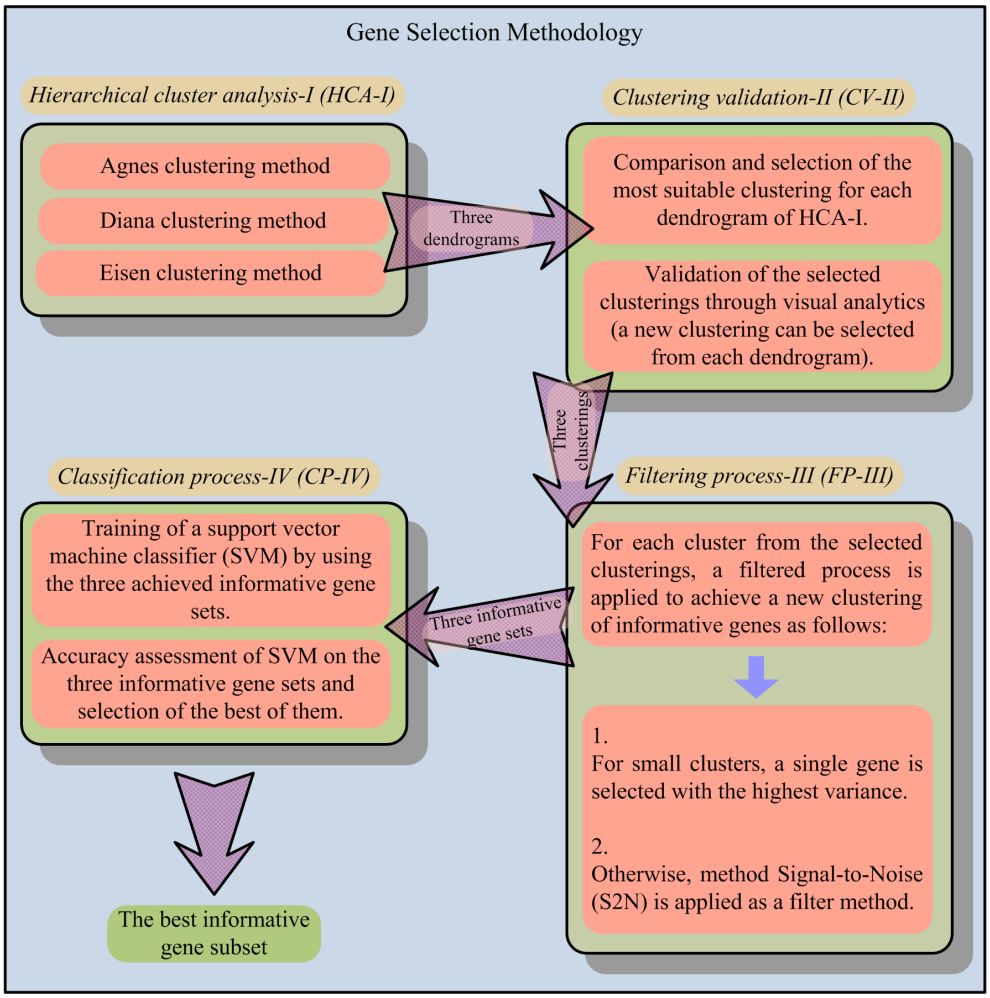

Figure 1: Steps followed by the gene selection approach. There are four stages processing a dataset, HCA-I, CV-II, FP-III and CP-IV. Each stage links with the next one by converting the result of one at the input of another.

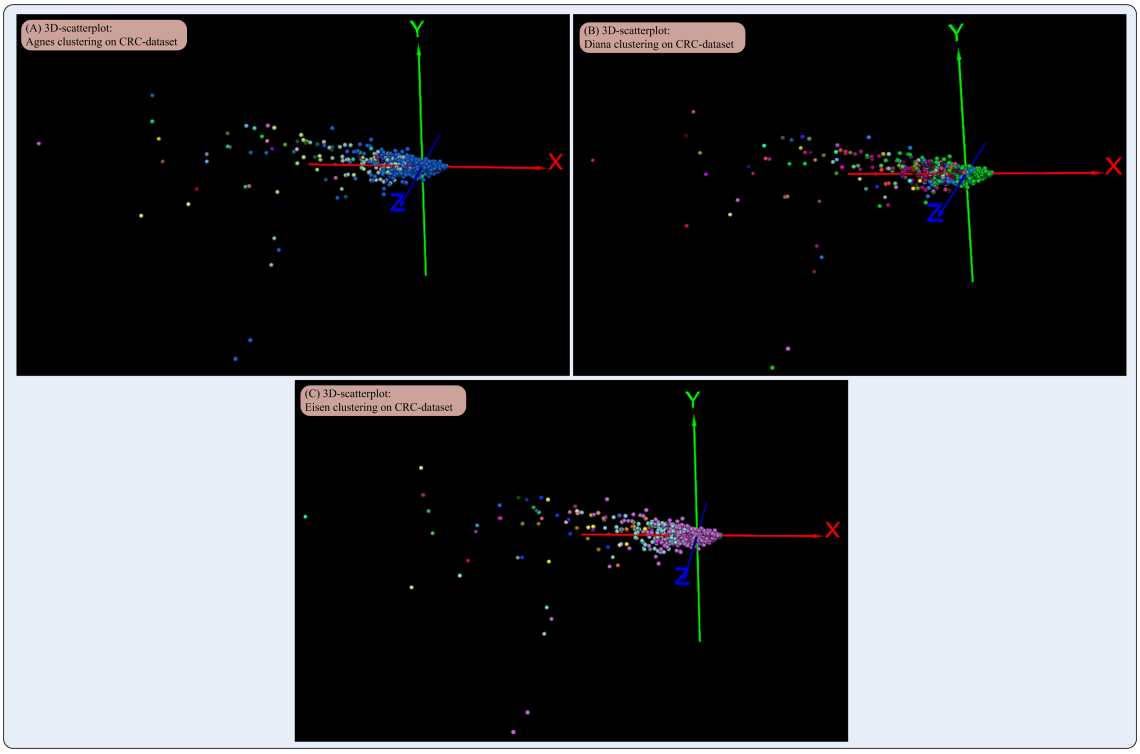

Figure 2: Three 3D-scatterplots of CRC-dataset representing each clustering given in Table 1. Gene-points belonging to the same cluster have the same colors.

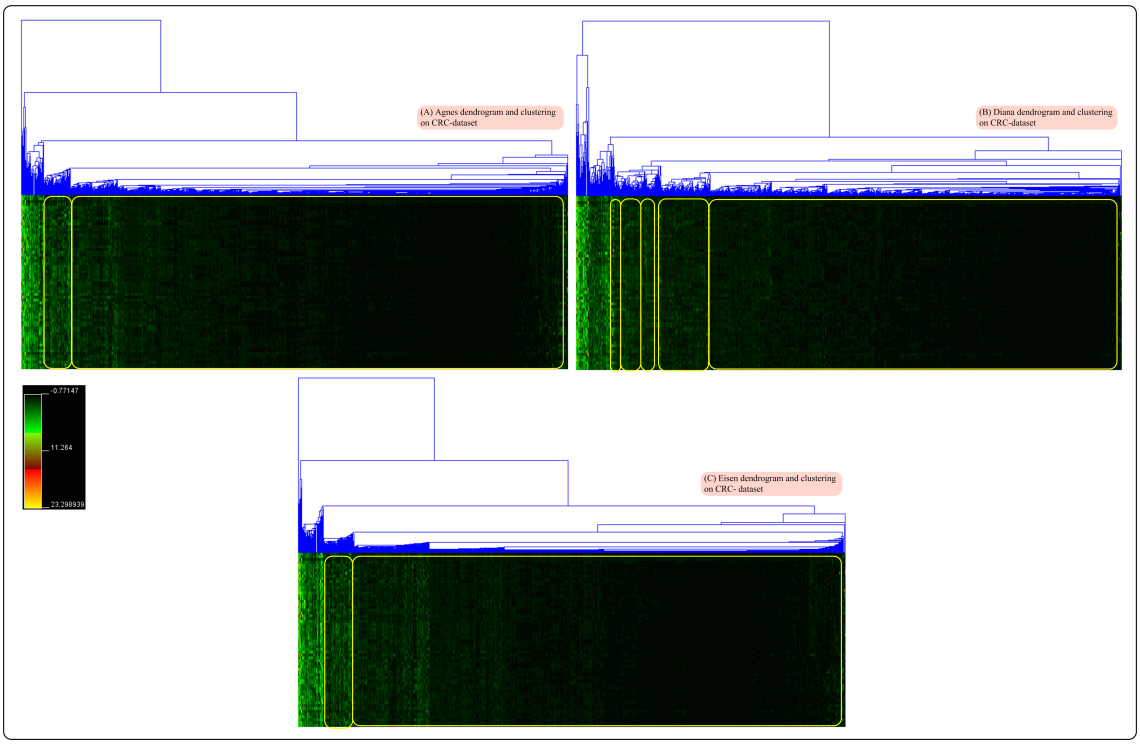

Figure 3: Three visualizations of dendrogram on heatmap representing each dendrogram and selected clustering in Table 1. Larger clusters of the level selected in each dendrogram have been stressed on the heatmap.