Supporting System for Detecting Pathologies

Carolina Zato, Juan F. De Paz, Fernando de la Prieta, and Beatriz Martín

Department of Computer Science and Automation, University of Salamanca
Plaza de la Merced s/n, 37008, Salamanca, Spain
{carol_zato,fcofds,fer,eureka}@usal.es

Abstract. Arrays CGH make possible the realization of tests on patients for the detection of mutations in chromosomal regions. Detecting these mutations allows to carry out diagnoses and to complete studies of sequencing in relevant regions of the DNA. The analysis process of arrays CGH requires the use of mechanisms that facilitate the data processing by specialized personnel since traditionally, a segmentation process is needed and starting from the segmented data, a visual analysis of the information is carried out for the selection of relevant segments. In this study a CBR system is presented as a supporting system for the extraction of relevant information in arrays CGH that facilitates the process of analysis and its interpretation.

Keywords: CGH arrays, knowledge extraction, visualization, CBR system.

1 Introduction

Arrays CGH (Comparative Genomic Hybridization) [39] are a type of microarrays that allows analyzing the information of the gains, losses and amplifications [36] in regions of the chromosomes for the detection of mutations. These types of microarrays unlike expression arrays do not measure the expression level of the genes; this is the reason why its use and analysis differ from the provided by expression arrays. The data obtained by the arrays CGH allows detecting automatically the mutations that characterize certain pathologies [29] [25]. Moreover, this information is useful to cross it with genetic sequencing, facilitating the analysis of the genetic sequencings and the sequencing tasks [6].

Microarray-based CGH and other large-scale genomic technologies are now routinely used to generate a vast amount of genomic profiles. Exploratory analysis of this data is crucial in helping to understand the data and to help form biological hypotheses. This step requires visualization of the data in a meaningful way to visualize the results and to perform first level analyses [32]. At present, tools and software already exist to analyze the data of arrays CGH, such as CGH-Explorer [24], ArrayCyGHt [19], CGHPRO [7], WebArray [38] or ArrayCGHbase [27], VAMP [32]. The problem of these tools is that they follow a static processing flow, without the possibility of storing or selecting those techniques that suit the samples of each case best. Therefore, these tools do not permit to personalize the flow of actions for the extraction of knowledge or to store preferences that can be useful in future processes with similar needs. The tool that is presented incorporates automatic procedures that can carry out

J. Cabestany, I. Rojas, and G. Joya (Eds.): IWANN 2011, Part II, LNCS 6692, pp. 669–676, 2011.

the analysis and the visual representations, facilitating the extraction of information with the most suitable processing flow. This allows the revision of the information by personnel without a great statistical knowledge and guarantees the obtaining of a better analysis automatically.

The process of arrays CGH analysis is decomposed in a group of structured stages, although most of the analysis process is done manually from the initial segmentation of the data. The initial data is segmented [35] to reduce the number of gains or losses fragments to be analyze. The segmentation process facilitates the later analysis of the data and is important to be able to represent a visualization of the data. Normally, the interpretation of the data is carried out manually from the visualization of the segmented data, however, when great amounts of these data have to be analyzed, it is necessary to create a decision support process.

For this reason, in this work a CBR system is included to facilitate the analysis and the automatic interpretation of the data by means of the configuration of analysis flows and the incorporation of flows based on predefined plans. The execution flows include procedures for the accomplishment of segmentation, classification, extraction of automatic information and visualization. The classification process facilitates the diagnosis of patients based on previous data; the process of knowledge extraction selects the differentiating regions of sets of patients by means of statistical techniques. Finally, the visualization process facilitates the revision of the results.

This article is divided as follows: section 2 describes the arrays CGH, section 3 describes our system, and section 4 presents the results and conclusions.

2 CBR-CGH System

CGH analysis allows the characterization of mutations that cause several cancers. The relationship between the chromosomal alterations and the prognosis of illness is well established. Recently, conventional array-based expression profiling has demonstrated that chromosomal alterations are associated with distinctive expression patterns. The system proposed in this work focuses on the detection of carcinogenic patterns in the data from CGH arrays, and is constructed from a CBR system that provides a classification and knowledge extraction technique based on previous cases.

The CBR developed system receives data from the analysis of chips and is responsible of establishing the workflow for classifying individuals based on evidence and existing data. The purpose of CBR is to solve new problems by adapting solutions that have been used to solve similar problems in the past [21]. The primary concept when working with CBRs is the concept of case. A case can be defined as a past experience, and is composed of three elements: a problem description which describes the initial problem, a solution which provides the sequence of actions carried out in order to solve the problem, and the final state which describes the state achieved once the solution was applied. The way cases are managed is known as the CBR cycle, and consists of four sequential steps which are recalled every time a problem needs to be solved: retrieve, reuse, revise and retain. Each of the steps of the CBR life cycle requires a model or method in order to perform its mission.

The algorithm selected for the retrieval of cases should be able to search the case base and selects the kind of default problems according to the analyzed data. In our

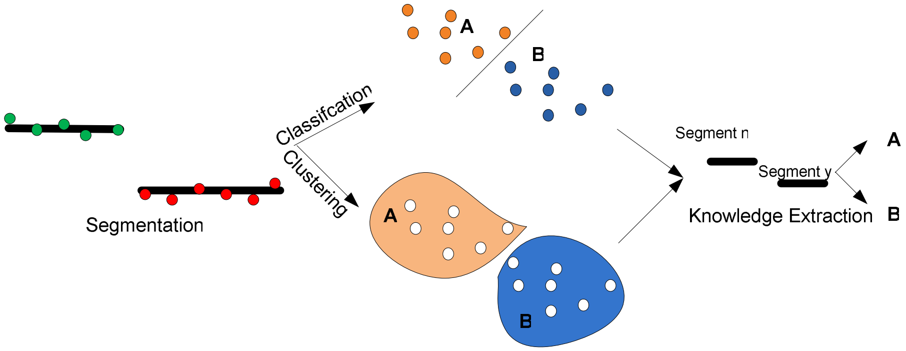

Fig. 1. Workflows in the classification, clustering, and knowledge extraction

case study, the system selects the workflows defined for each type of problem. The retrieved workflows are shown and the user selects one of them, then the activities are carried out. The revise phase consists of an expert revision for the proposed solution, and finally, the retain phase allows the system to learn from the experiences obtained in the three previous phases, consequently updating the cases memory.

The workflows set the sequence of actions in order to analyze the data. The kinds of default analysis are: clustering, classification and knowledge extraction. The figure 1 shows the available workflows and their activities since the initial state, for example a knowledge extraction process implies a segmentation and a clustering or classification activity.

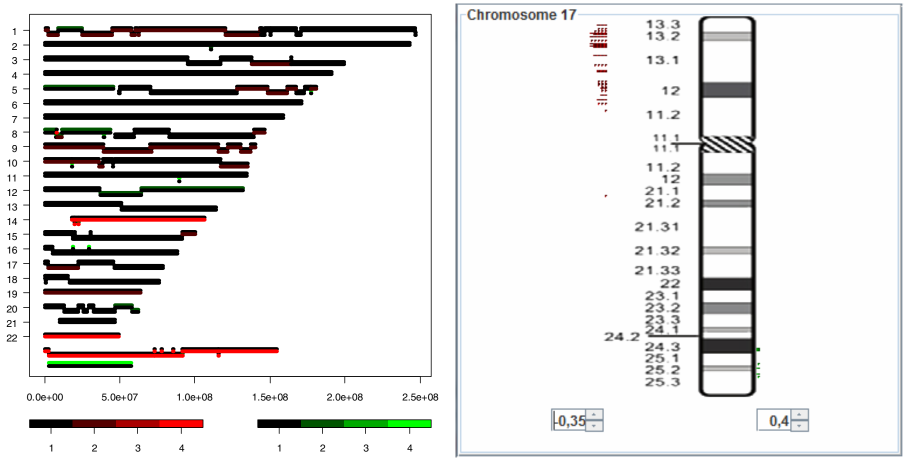

Fig. 2. Visualization of gains and losses using a) CGHcall and b) new method

In addition, a new visualization is provided to localized the mutations in an easier way, facilitating the identification of mutations that affects the gene codification among the large amount of genes. The figure 2a represents gains and losses using

CGHcall in R. The new visualization method is shown in the figure 2b, this visualization helps to locate the regions with mutations.

The system includes techniques for each of the activities (clustering, classification and knowledge extraction). Then, the applied algorithms in the steps are described.

2.1 Normalization and Segmentation

This stage constitutes the starting point for the treatment of the data and is necessary for the reduction of noise, the detection of losses and gains and the identification of breakpoints. The tool that is presented, through R Server, uses the package snapCGH [35], which allows both normalization and segmentation. Currently, many different segmentation algorithms are available, because of this, snapCGH incorporates software wrappers for several of these algorithms such as aCGH, DNACopy, GLAD and tilingArray. In [37][15] some comparisons between them can be found. The election of this package is due to the great acceptance, expansion and versatility, since it supplies many possibilities for the preprocessing.

2.2 Classification

The classification process is carried out according to a mixture of classifiers, although the system allows select a technique instead the mixture. A mixture of experts provide advances capacities by fusing the outputs of various processes (experts) and obtain the response more suitable for the final value [23] [28]. Mixtures of experts are also commonly used for classification and are usually called ensemble [41]. Some examples are the Bagging [5] or Ada-Boosting [11] algorithms. The classification algorithms can be divided in: decision trees, decision rules, probabilistic models, fuzzy models, based on functions, ensemble. The system selects these algorithms for each kind of method: decision rules RIPPER [8], One-R [16], M5 [17], decision trees J48 [31], CART [4] (Classification and Regression Trees), probabilistic models naive Bayes [10], fuzzy models K-NN (K-Nearest Neighbors) [1] and finally ensemble such as Bagging [5] and Ada-Boosting [11].

In order to calculate the final output of the system, RBF networks are used [13] [34]. The k cases retrieved in the previous phase are used by the classifiers and RBFs network as a training group that allows adapting its configuration to the new problem encountered before generating the initial estimation. The system presented in this article has a RBF network for each of the set of individual. Each of the RBF networks has as inputs the outputs estimated by the classifiers evaluated for the individual.

2.3 Clustering

Clustering techniques are typically broken down into the following categories [30] hierarchical, which include dendrograms [33], AGNES [18], DIANA [18], Clara [18]; neural networks such as SOM [20] (Self-Organized Maps), NG [26] (Neural Gas), GCS [12] (Growing Cell Structure; methods based on minimizing objective functions, such as k-means [14] and PAM [18] (Partition around medoids); or probabilistic-based models such as EM [2] (Expectation-maximization) and FANNY [18].

The provided methods are: in hierarchical clustering dendrograms [33], minimizing objective functions k-means [14] and PAM (Partitioning Around Medoids) [18] and in neural network SOCADNN (Self Organized Cluster Automatic Detection Neural Network) [3]. En el trabajo [3] se han realizado estudios sobre diferentes métodos de cluster y las ventajas que proporciona.

Hierarchical methods such as dendrograms do not require a number of clusters up front since they use a graphical representation to determine the number. Partition based methods as k-means and PAM, which optimize specific objective functions, have the disadvantage of requiring the number of clusters up front. Methods that are either hierarchical or minimize specific objective functions present certain deficiencies when it comes to recognizing groupings of individuals. ANN can adapt to the data surface, although they usually require additional time to do so. The SOM [20], have variants of learning methods that base their behaviour on methods similar to the NG [26]. They create a mesh that is adjusted automatically to a specific area. The ART networks can be considered as an alternative. The major disadvantage of these networks is the selection of the monitoring parameter [2] to determine the number of clusters. Another disadvantage is that the knowledge extraction is more complicated than in mesh-based networks, so learning is less evident.

2.4 Knowledge Extraction

Some techniques of the section 0 such as decision trees or rules, Bayesian networks or even rough sets could be applied in order to explain clusters or classifications although, the main objective in these problems is find maximum quantity of mutations that characterize a pathology. This information can be used in other studies as the sequencing of the concrete interesting regions with mutations. For this reason, statistical techniques are introduced in these activity for selecting the relevant segments. The introduced statistical techniques are broken down in non parametrics Kruskal-Wallis [42] and Mann-Whitney U-test [40] and parametrics ANOVA [9].

3 Results and Conclusions

In order to analyze the operation of the system, different data types of cancer, obtained from the data of the array CGH, were selected. In this case study we have 43 patients with GIST cancer, the data contain 4 kinds of pathologies: KIT Mutants, Wild-type, PDGFRA Mutants and BRAF Mutant, the pathology BRAF was removed because there was just one case with this illness. These data were previously classified, since the knowledge extraction is carried out from the previous classification. The data contain for each patients the kind of GIST and the segments with the gains and losses. The result of the relevant regions is shown in Table 1. Kruskal Wallis was applied for the extraction of this information, since the variables did not follow a normal distribution and therefore, a non-parametric test was required.

The figure 3 shows the highlighted region in the table 1. This region presents relevant differences among the detected GIST. In the box plots of the figure 3, PDGFRA doesn’t have losses or it presents gains in the region where the others present losses or they don't have variations. We can validate the others regions in similar way.

Table 1. Total number of hits for the different classifiers

Chromosome	Start	End	Nclone	Wide
8	139136846	146250764	314	7113918
15	30686790	91341204	2425	60654414
23	91485305	91537583	3	52278
22	134661	49565815	1491	49431154
20	58058472	62363573	200	4305101
8	39535654	43647062	143	4111408
8	7789936	8132138	3	342202
8	11665221	39341523	879	27676302
3	137653537	163941171	784	26287634
15	56257	18741715	15	18685458
1	**9110683**	**24996793**	**548**	**15886110**
9	70803414	70803414	9	146631
20	47048133	58039998	342	10991865
15	20249885	30298095	302	10048210

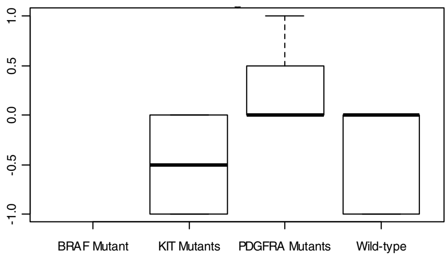

Fig. 3. Box plot for the region 9110683, 24996793

Although the system is still in a development phase, it is able to detect variations that allow characterizing different pathologies automatically. In addition, it permits the redefinition of execution flows, storing the sequence of actions that previously were considered satisfactory for its later use.

Acknowledgements. This work has been supported by the MICINN TIN 2009-13839-C03-03.

References

[1] Aha, D., Kibler, D., Albert, M.K.: Instance-based learning algorithms. Machine Learning 6, 37–66 (1991)

[2] Akhbardeh, A., Nikhil, Koskinenb, P.E., Yli-Harja, O.: Towards the experimental evaluation of novel supervised fuzzy adaptive resonance theory for pattern classification. Pattern Recognition Letters 29(8), 1082–1093 (2008)

[3] Bajo, J., De Paz, J.F., Rodríguez, S., González, A.: A new clustering algorithm applying a hierarchical method neural network. Logic Journal of IGPL (in Press)

[4] Breiman, L., Fried, J.H., Olshen, R.A., Stone, C.J.: Classification and regression trees. Wadsworth International Group, Belmont (1984)

[5] Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1984)

[6] Brown, P.O., Botstein, D.: Exploring the new world of the genome with DNA microarrays. Nature Genetics 21, 33–37 (1999)

[7] Chen, W., Erdogan, F., Ropers, H., Lenzner, S., Ullmann, R.: CGHPRO- a comprehensive data analysis tool for array CGH. BMC Bioinformatics 6(85), 299–303 (2005)

[8] Cohen, W.W.: Fast effective rule induction. In: Proceedings of the 12th International Conference on Machine Learning, pp. 115–123. Morgan Kaufmann, San Francisco (1995)

[9] De Haan, J.R., Bauerschmidt, S., van Schaik, R.C., Piek, E., Buydens, L.M.C., Wehrens, R.: Robust ANOVA for microarray data. Chemometrics and Intelligent Laboratory Systems 98(1), 38–44 (2009)

[10] Duda, R.O., Hart, P.: Pattern classification and Scene Analysis. John Wisley & Sons, New York (1973)

[11] Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Thirteenth International Conference on Machine Learning, pp. 148–156 (1996)

[12] Fritzke, B.: A growing neural gas network learns topologies. Advances in Neural Information Processing Systems 7, 625–632 (1995)

[13] Fritzke, B.: Fast Learning with Incremental RBF Networks. Neural Processing Letters 1(1), 2–5 (1994)

[14] Hartigan, J.A., Wong, M.A.: A K-means clustering algorithm. Applied Statistics 28, 100–108 (1979)

[15] Hofmann, W.A., Weigmann, A., Tauscher, M., Skawran, B., Focken, T., Buurman, R., Wingen, L.U., Schlegelberger, B., Steinemann, D.: Analysis of Array-CGH Data Using the R and Bioconductor Software Suite. Comparative and Functional Genomics, Article ID 201325 (2009)

[16] Holmes, G., Hall, M., Prank, E.: Generating Rule Sets from Model Trees. In: Advanced Topics in Artificial Intelligence, vol. 1747/1999, pp. 1–12 (2007)

[17] Holte, R.C.: Very simple classification rules perform well on most commonly used datasets. Machine Learning 11, 63–91 (1993)

[18] Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)

[19] Kim, S.Y., Nam, S.W., Lee, S.H., Park, W.S., Yoo, N.J., Lee, J.Y., Chung, Y.J.: Array-CyGHt, a web application for analysis and visualization of array-CGH data. Bioinformatics 21(10), 2554–2555 (2005)

[20] Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics, 59–69 (1982)

[21] Kolodner, J.: Case-Based Reasoning. Morgan Kaufmann, San Francisco (1993)

[22] Brunelli, R.: Histogram Analysis for Image Retrieval. Pattern Recognition 34, 1625–1637 (2001)

[23] Lima, C.A.M., Coelho, A.L.V., Von Zuben, F.J.: Hybridizing mixtures of experts with support vector machines: Investigation into nonlinear dynamic systems identification. Information Sciences 177(10), 2049–2074 (2007)

[24] Lingjaerde, O.C., Baumbush, L.O., Liestol, K., Glad, I.K., Borresen-Dale, A.L.: CGH-explorer, a program for analysis of array-CGH data. Bioinformatics 21(6), 821–822 (2005)

[25] Mantripragada, K.K., Buckley, P.G., Diaz de Stahl, T., Dumanski, J.P.: Genomic microarrays in the spotlight. Trends Genetics 20(2), 87–94 (2004)

[26] Martinetz, T., Schulten, K.: A neural-gas network learns topologies. Artificial Neural Networks 1, 397–402 (1991)

[27] Menten, B., Pattyn, F., De Preter, K., Robbrecht, P., Michels, E., Buysse, K., Mortier, G., De Paepe, A., van Vooren, S., Vermeesh, J., et al.: ArrayCGHbase: an analysis platform for comparative genomic hybridization microarrays. BMC Bioinformatics 6(124), 179–187 (2006)

[28] Nguyena, M.H., Abbassa, H.A., Mckay, R.I.: A novel mixture of experts model based on cooperative coevolution. Neurocomputing 70, 155–163 (2006)

[29] Pinkel, D., Albertson, D.G.: Array comparative genomic hybridization and its applications in cancer. Nature Genetics 37, 11–17 (2005)

[30] Po, R.W., Guh, Y.Y., Yang, M.S.: A new clustering approach using data envelopment analysis. European Journal of Operational Research 199(1), 276–284 (2009)

[31] Quinlan, J.R.: C4.5: Programs For Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco (1993)

[32] Rosa, P., Viara, E., Hupé, P., Pierron, G., Liva, S., Neuvial, P., Brito, I., Lair, S., Servant, N., Robine, N., Manié, E., Brennetot, C., Janoueix-Lerosey, I., Raynal, V., Gruel, N., Rouveirol, C., Stransky, N., Stern, M., Delattre, O., Aurias, A., Radvanyi, F., Barillot, E.: VAMP: Visualization and analysis of array-CGH, transcriptome and other molecular profiles Bioinformatics. Bioinformatics 22(17), 2066–2073 (2006)

[33] Saitou, N., Nie, M.: The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425 (1987)

[34] Savitha, R., Suresh, S., Sundararajan, N.: A fully complex-valued radial basis function network and its learning algorithm. Int. Journal of Neural Systems. 19(4), 253–267 (2009)

[35] Smith, M.L., Marioni, J.C., Hardcastle, T.J., Thorne, N.P.: snapCGH: Segmentation, Normalization and Processing of aCGH Data Users' Guide. Bioconductor (2006)

[36] Wang, P., Young, K., Pollack, J., Narasimham, B., Tibshirani, R.: A method for callong gains and losses in array CGH data. Biostat. 6(1), 45–58 (2005)

[37] Willenbrock, H., Fridlyand, J.: A comparison study: applying segmentation to array CGH data for downstream analyses. Bioinformatics 21(22), 4084–4091 (2005)

[38] Xia, X., McClelland, M., Wang, Y.: WebArray, an online platform for microarray data analysis. BMC Bionformatics 6(306), 1737–1745 (2005)

[39] Ylstra, B., Van den Ijssel, P., Carvalho, B., Meijer, G.: BAC to the future! or oligonucleotides: a perspective for microarray comparative genomic hybridization (array CGH). Nucleic Acids Research 34, 445–450 (2006)

[40] Yue, S., Wang, C.: The influence of serial correlation on the Mann-Whitney test for detecting a shift in median. Advances in Water Resources 25(3), 325–333 (2002)

[41] Zhanga, H., Lu, J.: Creating ensembles of classifiers via fuzzy clustering and deflection. Fuzzy Sets and Systems 161(13), 1790–1802 (2010)

[42] Kruskal, W., Wallis, W.: Use of ranks in one-criterion variance analysis. Journal of American Statistics Association (1952)