Visual Analysis Tool in Comparative Genomics

Juan F. De Paz, Carolina Zato, María Abáigar, Ana Rodríguez-Vicente, Rocío Benito, and Jesús M. Hernández

Abstract. Detecting regions with mutations associated with different pathologies is an important step in selecting relevant genes, proteins or diseases. The corresponding information of the mutations and genes is distributed in different public sources and databases, so it is necessary to use systems that can contrast different sources and select conspicuous information. This work presents a visual analysis tool that automatically selects relevant segments and the associated genes or proteins that could determine different pathologies.

1 Introduction

Different techniques presently exist for the analysis and identification of pathologies at a genetic level. Along with massive sequencing, which allows the exhaustive study of mutations, the use of microarrays is highly extended. With regards to microarrays, there are two primary types of chips according to the direction of the analysis that will be carried out: expression arrays and array CGH (Comparative Genomic Hybridization) [18]. CGH arrays (aCGH) are a type of microarray that can analyze information on the gains, losses and amplifications [16] in regions of the chromosomes to detect mutations that can determine some pathologies [13] [11].

Juan F. De Paz · Carolina Zato
Department of Computer Science and Automation, University of Salamanca
Plaza de la Merced, s/n, 37008, Salamanca, Spain
e-mail: {fcofds,carol_zato}@usal.es

María Abáigar · Ana Rodríguez-Vicente · Rocío Benito · Jesús M. Hernández
IBMCC, Cancer Research Center, University of Salamanca-CSIC, Spain
e-mail: {mymary,anita82,beniroc,jhmr}@usal.es

Jesús M. Hernández
Servicio de Hematología, Hospital Universitario de Salamanca, Spain

M.P. Rocha et al. (Eds.): 6th International Conference on PACBB, AISC 154, pp. 121–127.
springerlink.com

Microarray-based CGH and other large-scale genomic technologies are now routinely used to generate a vast amount of genomic profiles. An exploratory analysis of this data is critical in helping to understand the data and form biological hypotheses. This step requires visualization of the data in a meaningful way to visualize the results and to perform first level analyses [15]. At present, tools and software already exist to analyze the data of arrays CGH, such as CGH-Explorer [10], ArrayCyGHt [9], CGHPRO [3], WebArray [17] or ArrayCGHbase [12], VAMP [15]. The problem with these tools is the lack of usability and of an interactive model. For this reason, it is necessary to create a visual tool to analyse the data in a simpler way.

The process of arrays CGH analysis is broken down into a group of structured stages, although most of the analysis process is done manually from the initial segmentation of the data. Once the segmentation is finished, the next step is to perform a visual analysis of the data using different tools, which is a quite slow process. For this reason, the system tries to facilitate the analysis and the automatic interpretation of the data by selecting the relevant genes, proteins and information from the previous classification of pathologies. The system provides several representations in order to facilitate the visual analysis of the data. The information for the identified genes, CNV, pathologies etc. is obtained from public databases.

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

2 aCGH Analysis Tool

aCGH is a technique that can detect copy number variations in patients who have undergone different mutations in chromosomic regions. Typically, the variations have been previously catalogued, allowing the existing information to be used to catalogue and evaluate the mutation. In this case study, the cases are defined according to the segments into which the chromosomic regions have been fragmented.

The developed system receives data from the analysis of chips and is responsible for representing the data for extracting relevant segments on evidence and existing data. Working from the relevant cases, the first step consists of selecting the information about the genes and transcripts stored in the databases. This information will be associated to each of the segments, making it possible to quickly consult the data and reveal the detected alterations at a glance. The data analysis can be carried out automatically or manually.

2.1 Automatic Analysis

Knowledge extraction algorithms can be categorized as decision trees, decision rules, probabilistic models, fuzzy models, based on functions, statistics, or gain

functions. Some of these algorithms include: decision rules RIPPER [4], One-R [7], M5 [8], decision trees J48 [14], CART [2] (Classification and Regression Trees), probabilistic models naive Bayes [6], fuzzy models K-NN (K-Nearest Neighbors) [1] and finally statistical techniques, such as non parametrics Kruskal-Wallis [21] and Mann-Whitney U-test [19] for two groups, and parametrics Chi Squared [22], ANOVA [5]. The gain functions are a particular case of the techniques used in decision trees and decision rules for selecting the attributes, which is why they are not considered separately.

For this particular system, the use of decision trees was chosen to select the main genes of the most important pathologies, specifically J48 [14] in its implementation for Weka [20]. However, if the system needs a generic selection, the gain functions are chosen (specifically, Chi Squared [22], which is also implemented in the Weka library). Chi Squared was chosen because it is the technique that makes it possible to work with different qualitative nominal variables to study factor and its response. The contrast of Chi Squared makes it possible to obtain as output the values that can sort the attributes by their importance, providing an easier way to select the elements. As an alternative, gain functions could be applied in decision trees, providing similar results.

2.2 Visual Analysis

A visual analysis is performed of the data provided by the system and the information recovered from the databases. New visualizations are performed in order to more easily locate the mutations, thus facilitating the identification of mutations that affect the codification of genes among the large amount of genes. Visualization facilitates the validation of the results due to the interactivity and ease of use of previous information. Existing packages such as CGHcall [23] in R do not display the results in an intuitive way because it is not possible to associate segments with regions and they do not allow interactivity.

The system provides a visualization to select the regions with more variants and relevant regions in different pathologies. The visualizations make is possible to extract information from databases using a local database.

2.3 Reviewing Process

Once the relevant segments have been selected, the researchers can introduce information for each of the variants. The information is stored in a local database. These data are considered in future analyses although they have to be reviewed in detail and contrasted by the scientific community. The information is shown in future analyses with the information for the gains and losses. However, because only the information from public databases is considered reliable, this information is not included in the reports.

3 Results and Conclusions

The system was applied to two different kinds of CGH arrays: BAC aCGH, and Oligo aCGH. The information obtained from the BAC aCGH after segmenting and normalizing is represented in table 1. As shown in the figure, there is one patient for each column. The rows contain the segments so that all patients have the same segments. Each segment is a tuple composed of three elements: chromosome, initial region and final region. The values v_{ij} represent gains and losses for segment i and patient j. If the value is positive, or greater than the threshold, it is considered a gain; if it is lower than the value, it is considered a loss.

Table 1 BAC aCGH normalized and segmented

Segment	Patient 1	Patient 2	...	Pantient n
Init-end	v_{11}	v_{12}	...	v_{13}
Init-end	v_{21}	v_{22}	...	v_{23}

The system includes the databases because it extracts the information from genes, proteins and diseases. These databases have different formats but basically there is a tuple of three elements for each row (chromosome, start, end, other information). Altogether, the files downloaded from UCSC included slightly more than 70,000 registries

Figure 1 displays the information for 18 oligo arrays cases. Only the information corresponding to chromosome 5 is shown. The green lines represent gains for the patient in the associated region of the chromosome, while the red lines represent losses. The user can draw the CNVs and use this information to select the relevant information. For example, in figure 3 the CNVs are represented in blue. We can see that there is a gain region with a high incidence in the individuals but this region is not relevant because it belongs to a CNV.

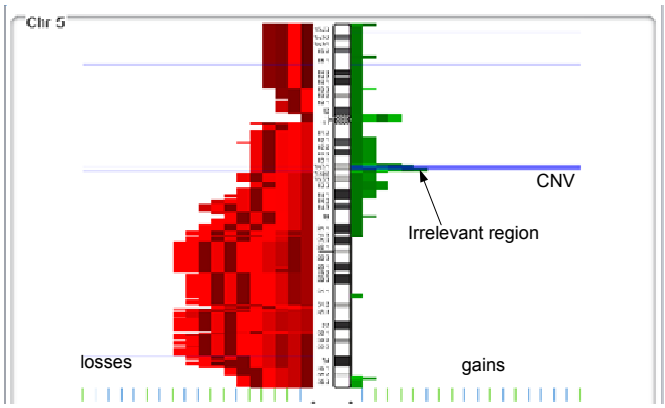

Fig. 1 Selection of segments and genes automatically

When performing the visual analysis, users can retrieve information from a local database or they can browse through UCSC. For example, figure 2 contains the information for the segment belonging to the irrelevant region shown in the previous image.

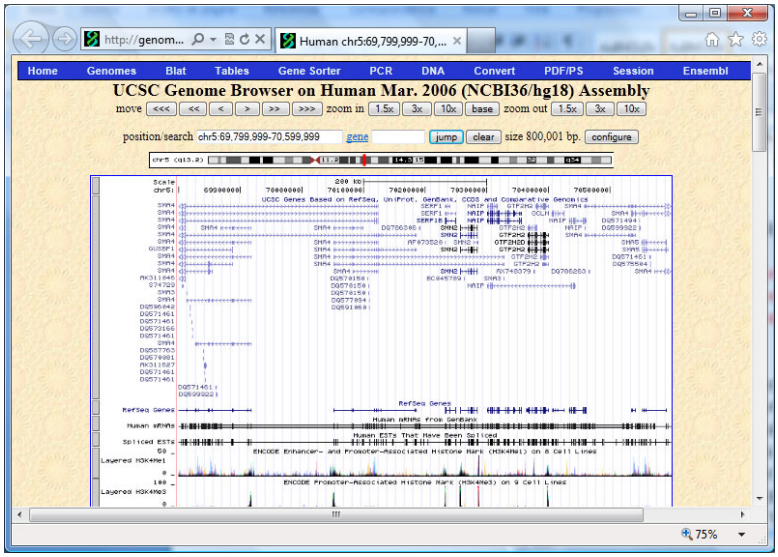

Fig. 2 Browse through UCSC

In order to facilitate the revision and learning phases for the expert, a different visualization of the data is provided. This view helps to verify the results obtained by the hypothesis contrast regarding the significance of the differences between pathologies. Figure 3 shows a bar graph where one bar represents each individual and is divided into different segments with an amplitude proportional to the width of the segment gain (green) or loss (red). We can see that the blue individuals (rectangle over the bars) are not in the range of the green individuals because they remain deactivated when we select the green individuals.

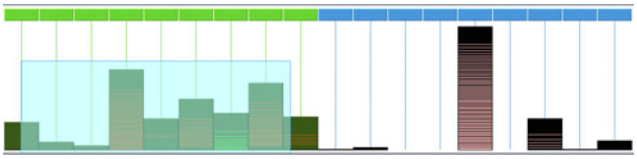

Fig. 3 Selection of segments and genes automatically

The presented system facilitates the use of different sources of information to analyze the relevance in variations located in chromosomic regions. The system is able to select the genes, variants, genomic duplications that characterize pathologies automatically, using several databases. This system allows the management of external sources of information to generate final results. The provided visualizations make it possible to validate the results obtained by an expert more quickly and easily

Acknowledgments. 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] Breiman, L., Fried, J.H., Olshen, R.A., Stone, C.J.: Classification and regression trees. Wadsworth International Group. (1984)
[3] 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)
[4] 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)
[5] 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)
[6] Duda, R.O., Hart, P.: Pattern classification and Scene Analysis. John Wisley & Sons, New York (1973)
[7] Holmes, G., Hall, M., Prank, E.: Generating Rule Sets from Model Trees. Advanced Topics in Artificial Intelligence 1747(1999), 1–12 (2007)
[8] Holte, R.C.: Very simple classification rules perform well on most commonly used datasets. Machine Learning 11, 63–91 (1993)
[9] Kim, S.Y., Nam, S.W., Lee, S.H., Park, W.S., Yoo, N.J., Lee, J.Y., Chung, Y.J.: ArrayCyGHt, a web application for analysis and visualization of array-CGH data. Bioinformatics 21(10), 2554–2555 (2005)
[10] 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)
[11] 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)
[12] 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)
[13] Pinkel, D., Albertson, D.G.: Array comparative genomic hybridization and its applications in cancer. Nature Genetics 37, 11–17 (2005)
[14] Quinlan, J.R.: C4.5: Programs For Machine Learning. Morgan Kaufmann Publishers Inc. (1993)

[15] 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.: Visualization and analysis of array-CGH, transcriptome and other molecular profiles. Bioinformatics 22(17), 2066–2073 (2006)

[16] 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)

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

[18] 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)

[19] 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)

[20] `http://www.cs.waikato.ac.nz/ml/weka/`

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

[22] Kenney, J.F., Keeping, E.S.: Mathematics of Statistics, Pt. 2, 2nd edn. Van Nostrand, Princeton (1951)

[23] Van de Wiel, M.A., Kim, K.I., Vosse, S.J., Van Wieringen, W.N., Wilting, S.M., Ylstra, B.: CGHcall: calling aberrations for array CGH tumor profiles. Bioinformatics 23(7), 892–894 (2007)