Comparative Genomics with Multi-agent Systems

Juan F. De Paz, Carolina Zato, Fernando de la Prieta, Javier Bajo,
Juan M. Corchado, and Jesús M. Hernández

Abstract. The detection of the regions with mutations associated with different pathologies is an important step for selecting relevant genes. 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 proposes a virtual organization of agents that can analyze and interpret the results from Array-based comparative genomic hybridization, thus facilitating the traditionally manual process of the analysis and interpretation of results.

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

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. CGH arrays (aCGH) (Array-based comparative genomic hybridization) are a type of microarray that can analyze information on the gains, losses and amplifications [7] in regions of the chromosomes to detect mutations [5], [3]. Expression arrays measure the expression level of the genes. aCGH are currently used to detect relevant regions that may require deeper analysis. In these cases, it is necessary to

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

Jesús M. Hernández
IBMCC, Cancer Research Center, University of Salamanca-CSIC, Spain
e-mail: jhmr@usal.es

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

A. Abraham and S.M. Thampi (Eds.): Intelligent Informatics, AISC 182, pp. 175–181.
springerlink.com

work with vast amounts of information, which necessitates the creation of a system that can facilitate the automatic analysis of data that, in turn, facilitates the extraction of relevant information using different data bases. For this reason, it is necessary to automate the aCGH processing.

aCGH, also called microarray analysis, is a new cytogenetic technology that evaluates areas of the human genome for gains or losses of chromosome segments at a higher resolution than traditional karyotyping. When working with aCGH, segments of DNA (Deoxyribonucleic Acid) are selected from public genome databases based upon their location in the genome. Computer software analyzes the fluorescent signals for areas of unequal hybridization of patient versus control DNA, signifying a DNA dosage alteration (deletion or duplication). These arrays offer genome-covering resolution that can offer precise delineation of breakpoints. This is important in determining common regions of overlap and implicated genes. Due to their small target size, oligonucleotide arrays suffer from poorer signal to noise ratios that often results in a significant number of false-positive outliers. At present, tools and software already exist to analyze the data of arrays CGH, such as CGH-Explorer [2], ArrayCyGHt [12], CGHPRO [1], WebArray [8] or ArrayCGHbase [4], VAMP [6]. 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. This study presents a multi-agent system [10] that defines roles to automatically perform the different stages of the analysis. In the first stage, the data are segmented [11] to reduce the number of gains or losses fragments to be analyzed. The following steps vary in terms of the type of analysis being performed and include: grouping, classification, visualization, or extraction of information from different sources. 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, CNVs (Copy-number variations), 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 Multi-agent System

The multi-agent system designed to analyze our data is general enough that it can be adapted for other types of data analysis. The multi-agent system is divided into different layers: the analysis layer, the information management layer. 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 Analysis Roles

The analysis roles contains the agents responsible for performing the actual microarray analyses. The information management layer compiles the information from the database and generates local databases to facilitate their analysis. The visualization layer facilitates the management of both the information and the algorithms; it displays the information and the results obtained after applying the existing algorithms at the analysis layer.

The agents at the analysis layer adapt to the specific class of microarray, in this case the aCGH, and within the aCGH they adapt to the different types of microarrays with which they work. To perform the data analysis, the agents are incorporated for: segmentation, Knowledge extraction, and Clustering.

The segmentation process is performed by taking into account the differential normalization for gains and losses. The segmentation process is based on the mad1dr (median absolute deviation, 1st derivative) value for each of the arrays, which determines the threshold for gains or losses that is considered relevant for each case. This metric provides a surrogate measure of experimental noise.

For this particular system, the use of chi Square 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 Square 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 Information Management Roles

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.

Besides the system incorporates a role to retrieve information from UCSC (University of California Santa Cruz) and use this information to generate reports. This information is important in order to select the relevant segments.

3 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 [9] 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.

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 [9] 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.

4 Results and Conclusions

In order to analyze the operation of the system, different data types of array CGH were selected. 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 Tab. 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_{1n}
Init-end	v_{21}	v_{22}	...	v_{2n}

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.

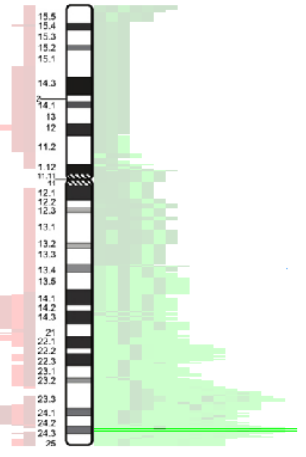

Fig. 1 Selection of segments and genes automatically

Figure 1 displays the information for 18 oligo arrays cases. Only the information corresponding to chromosome 11 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 select the regions and use these highlighted regions to generate reports.

When performing the visual analysis, users can retrieve information from a local database or they can browse through UCSC. For example, figure 2 contains a

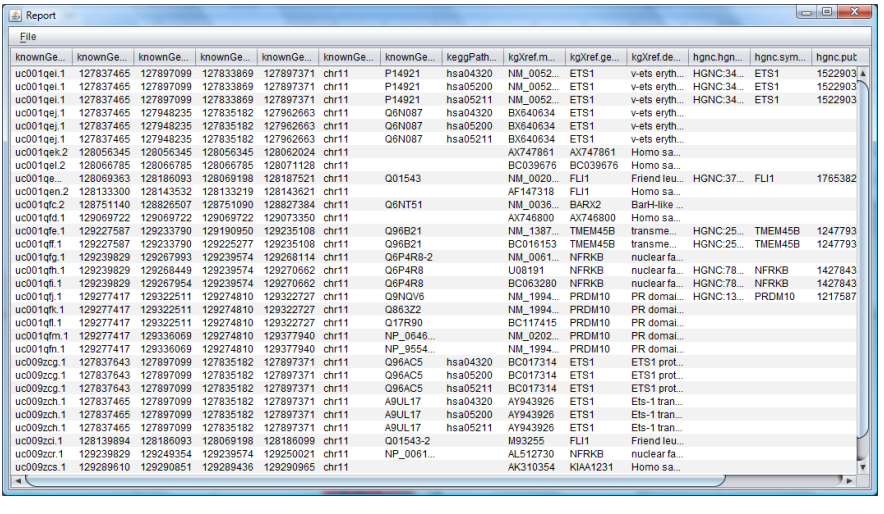

Fig. 2 Report with relevant genes

report with the information for the segment belonging to the irrelevant region shown in the previous image.

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 dendrogram with the information of the groups. The expert can review the clusters and modify the group belong each patient selecting each patient.

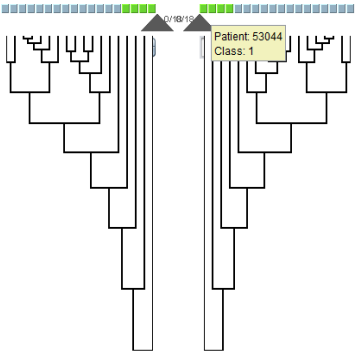

Fig. 3 Reviewing clustering process

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] 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)

[2] 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)

[3] 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)

[4] 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.: Array CGH base: an analysis platform for comparative genomic hybridization microarrays. BMC Bioinformatics 6(124), 179–187 (2006)

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

[6] 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 22(17), 2066–2073 (2006)

[7] 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)

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

[9] 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)

[10] Argente, E., Botti, V., Carrascosa, C., Giret, A., Julian, V., Rebollo, M.: An abstract architecture for virtual organizations: The THOMAS approach. Knowledge and Information Systems 29(2), 379–403 (2011)

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

[12] 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)