ANALYSIS OF ACGH INTEGRATING DIFFERENT SOURCES OF INFORMATION BY MEANS OF A CBR

Carolina Zato
Departamento Informática y Automática. Universidad de Salamanca
Plaza de la Merced s/n, 37008, Salamanca, Spain
carol_zato@usal.es

Juan F. De Paz
Departamento Informática y Automática. Universidad de Salamanca
Plaza de la Merced s/n, 37008, Salamanca, Spain
fcofds@usal.es

Javier Bajo
Facultad de Informática. Universidad Pontificia de Salamanca
Compañía 5, 37002, Salamanca, Spain
jbajope@upsa.es

Juan M. Corchado
Departamento Informática y Automática. Universidad de Salamanca
Plaza de la Merced s/n, 37008, Salamanca, Spain
corchado@usal.es

Abstract. Knowledge management is a key element in medical environments. Arrays CGH make possible the realization of tests on patients for the detection of mutations in chromosomal regions. The detection of the regions with mutations associated to different pathologies is an important step for the selection of 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 the use of systems that allow to contrast different sources for the selection of outstanding information. In this work, a case-based reasoning (CBR) system is presented to carry out the automatic selection of relevant segments and the associated genes or proteins that could determine different pathologies. The CBR system integrates statistical techniques for the selection of relevant genes and visualization techniques for the interpretation of the final results.

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

1. Introduction

Knowledge Management is a fundamental asset for businesses in the contemporary economy. Knowledge takes into account the organization of the businesses, individuals and the information (Takeishi, A. 2002). Knowledge management can be applied to different organizations and different contexts. Nowadays, knowledge management in medical contexts is acquiring a growing relevance. More specifically, knowledge management is specially relevant to detect and predict diseases. This paper focuses on knowledge management in medical environments, aimed at the identification and prediction of cancer using the information available from CGH arrays.

At present different techniques exists for the analysis and identification of pathologies at a genetic level. Along with the massive sequencing, that allows the exhaustive study of mutations, the use of microarrays is highly extended. Related with the microarrays, there are different types of chips according to the direction of the analysis to carry out, mainly the expression arrays and arrays CGH (Comparative Genomic Hybridization) (Ylstra, B. et al. 2006). Arrays CGH (aCGH) are a type of microarrays that allows analyzing the information of the gains, losses and amplifications (Wang, P. et al. 2005) in regions of the chromosomes for the detection of mutations that can determine some pathologies (Pinkel, D. & Albertson, D.G. 2005) (Mantripragada, K.K. et al. 2004). The expression arrays measure the expression level of the genes. aCGH are currently used to detect relevant regions susceptible to be deeper analyzed, this information is taken into to account for sequencing these regions throughout expression arrays and sequencers (Brown, P.O. and Botstein, D. 1999). For this reason, it is necessary to automate the aCGH processing, simplifying the location of those interesting genes before carry out the sequencing. The information about the variants is managed using data from databases in order to obtain knowledge.

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 (Rosa, P. et al. 2006). At present, tools and software already exist to analyze the data of arrays CGH, such as CGH-Explorer (Lingjaerde, O.C. et al. 2004), ArrayCyGHt (Kim, S.Y. et al. 2005), CGHPRO (Chen, W. et al. 2005), WebArray (Xia, X. et al. 2005) or ArrayCGHbase (Menten, B. et al. 2006), VAMP (Rosa, P. et al. 2006). The problem of these tools is that do not execute an automatic analysis of the data and the own user is in charge to analyze the information and to decide the steps to follow to process the data. For this reason, it is necessary to incorporate a process that helps to determine the interesting genes to be analyzed and the known transcripts for those genes in a simpler way.

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 (Smith, M.L. et al. 2006) 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. Once the segmentation is finished, the next step is the accomplishment of the visual analysis of the data using different tools, a quite slow process. For this reason, in this work a CBR system is included to facilitate

the analysis and the automatic interpretation of the data, selecting the relevant genes, proteins and relevant information for the previous classification of pathologies. The information of the identified genes is obtained from public databases. The CBR system is based on the CBR reasoning cycle, which will proceed selecting the relevant genes using knowledge extraction techniques like decision trees. 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. CGH Arrays

Array-based comparative genomic hybridization (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. Whereas traditional high-resolution chromosome analysis detects chromosome structure alterations at a resolution of 5 megabases (Mb) or greater, aCGH detects gains or losses of DNA, that cannot be seen by traditional karyotyping and may sometimes be only thousands of basepairs in size (Hixson, P. et al. 2006). aCGH has emerged as a powerful diagnostic technique for high resolution analysis of the human genome. It is a specific, sensitive, and rapid technique enabling detection of genomic arrangements and copy number changes. A variety of array CGH platforms are currently available, both commercially and in academic institutions. The choice of platform may depend on the type of data sought; however, the price, reproducibility, and standardization are crucial factors that need to be considered (Hixson, P. et al. 2006).

For the work with aCGH, segments of DNA are selected from public genome databases based upon their location in the genome. The clones are predominantly selected to target areas of the human genome that, when deleted or duplicated, are known or highly suspected to cause well-characterized genetic defects. Microarray printers attach the clones to a glass slide in an organized way to form a microarray. A typical microarray slide contains thousands of different clones representing targeted areas of the genome. Fluorescently labeled DNA from both patient and a known normal human control are applied to the slide and compete to attach or hybridize to their corresponding DNA segments. Computer software analyzes the fluorescent signals for areas of unequal hybridization of patient versus control DNA, signifying a DNA dosage alteration (deletion or duplication).

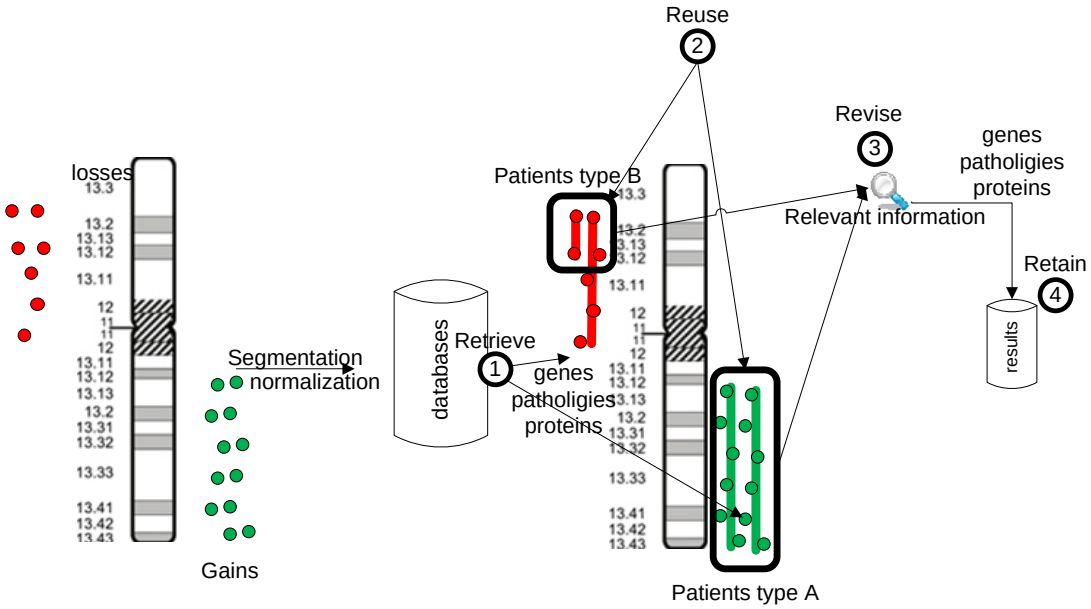
3. CBR-aCGH system

aCGH are a technique that allows to detect copy number variations in patients who undergo different mutations in chromosomic regions. Habitually, the variations are catalogued previously, reason why the existing information can be used to catalogue and to evaluate the mutation. In this case study, the cases are defined based on the segments in which the chromosomic regions have been fragmented. Therefore, in a CBR system, the retrieve and selection phase is adapted to get the most suitable information that solves the problem.

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 (Kolodner J. 1993). 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 genes and the known transcripts associated to gain or losses regions. The retrieved genes and relevant information's are shown with each of the segments to validate the obtained results. The revise phase consists of an expert revision for the proposed solution, and finally, the retain phase allows the system to stores the information considered relevant. The analysis process followed by the system is shown in figure 1. Next, the techniques applied during the different phases of the CBR cycle are described in detail.

Figure 1. Chromosome 19 losses in red and gains in green



3.1.Retrieve

The retrieve phase is split in two stages. In the first stage, the cases in relevance study are selected. The relevant cases are those segments that represent gains or losses. In the second stage, the processes of segmentation and normalization are carried out as described later. And finally, the information of the existing genes and transcripts is recovered from the existing information on the databases.

3.1.1. 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 (Smith, M.L. et al. 2006), 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 (Willenbrock, H. & Fridlyand, J. 2005) (Hofmann, W.A. et al. 2009) 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.

3.2.1. Retrieval of relevant information

Working from the relevant cases, the next step consists of selecting the stored information about the genes and transcripts in the databases. This information will be associated to each of the segments allowing a quick consult of the data and revealing the detected alterations at first sight.

3.2.Reuse

The knowledge extraction algorithms can be divided in: decision trees, decision rules, probabilistic models, fuzzy models, based on functions, statistics, gain functions. The system selects these algorithms for each kind of method: decision rules RIPPER (Cohen, W.W. 1995), One-R (Holmes, G. et al. 2007), M5 (Holte, R.C. 1993), decision trees J48 (Quinlan, J.R. 1993), CART (Breiman, L. et al. 1984) (Classification and Regression Trees), probabilistic models naive Bayes (Duda R.O. & Hart P. 1973), fuzzy models K-NN (K-Nearest Neighbors) (Aha D. et al. 1991) and finally statistical techniques as non parametrics Kruskal-Wallis (Kruskal, W., & Wallis. W. 1952) and Mann-Whitney U-test (Yue, S. & Wang, C. 2002) for two groups, and parametrics Chi Squared (Kenney, J.F. & Keeping, E.S. 1951), ANOVA (Cohen, W.W. 1995) The gain functions are a particular case of the techniques used in the decision trees and decision rules for the selection of the attributes, this is the reason why are not considered separately.

Particularly for this system, the use of decision trees has been chosen for the selection of the main genes of the most important pathologies, specifically J48 (Cohen, W.W. 1995) in its implementation for Weka. In a more general way, if the system needs a generic selection, the gain functions are chosen (concretely, Chi Squared (Kenney, J.F. & Keeping, E.S. 1951), which is also implemented in the Weka library). Chi Squared has been chosen because is the technique that allows working with different qualitative nominal variables to the studying factor and its response. The contrast of Chi Squared allows to obtained as output, values that can sort the attributes by its importance, providing a easier way to select the elements. As alternative, gain functions could be applied in decision trees, the results are similar.

3.3.Revise

The revision phase is carried out manually by a visual analysis of the data provided by the system and the information recovered from the databases. 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 visualization facilitates the validation of the results due to the interactivity and easy to use of previous information. Existing packages as CGHcall in R show the results in a little intuitive way because it is not possible to associate segments with regions and they don't allow interactivity.

3.4.Retain

Once the relevant segments have been selected, the information is stored. This information is not considered in future analysis because it has to be reviewed in detail and contrasted by the scientific community, then the information is included in public database and the information will be taken into account in future analysis. Only the information of public databases is considered reliable.

4. Results and conclusions

The system has been applied on two different kind of arrays CGH. The information of BAC aCGH after segmenting and normalizing is represented according to the figure 2. As we can see in the figure, there is a patient for each column. The rows contain the segments then 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 the segment i and patient j. If the value is positive greater than a threshold, it is considered as gain, and if is lower than a value it is considered a losses.

Figure 2. BAC aCGH normalized and segmented

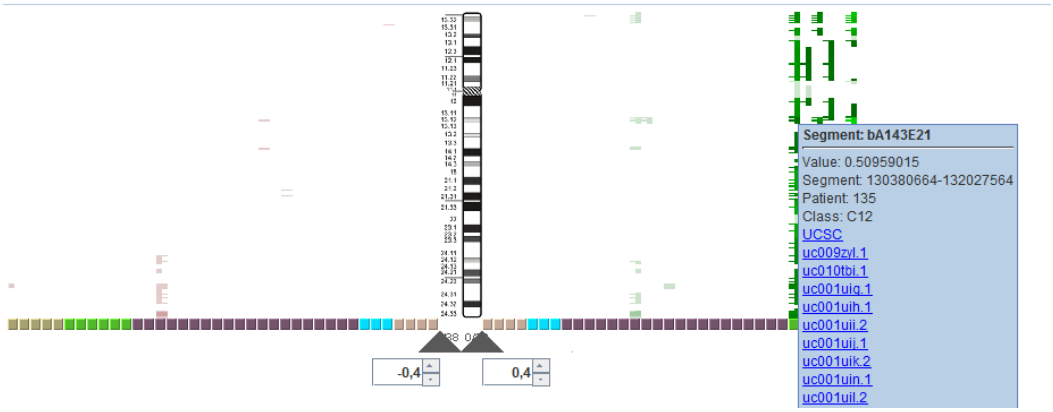
	Patient 1	Patient 2	Patient 3	...	Patient n
Segment 1 (Chromosome-init-end)	v_{11}	v_{12}	v_{13}	…	v_{1n}
Segment 2 (Chromosome-init-end)	v_{21}	v_{22}	v_{23}	…	v_{2n}
...	…	…	…	…	…
Segment m (Chromosome-init-end)	v_{m1}	v_{m2}	v_{m3}	…	v_{mn}

Finally, the system has the databases since the system extracts the information about genes, proteins and diseases. These databases have different format but basically it has a tuple of three elements for each gene (chromosome, start, end). Altogether, the file downloaded from UCSC counted with a little more than 70,000 registries

In figure 3 the information of the BAC arrays cases is shown with 38 cases with 5 different pathologies. Only the information corresponding with the chromosome 12 is shown. The green lines represents gains of the patient in the associated region of the chromosome, while the red lines represent losses. So, the figure shows that the green patients have gains while the rest of them present few variations. Automatically, the most relevant segments are highlighted as bright segments by the application of the hypothesis contrast Chi Squared. This technique facilitates the selection oh the relevant segments.

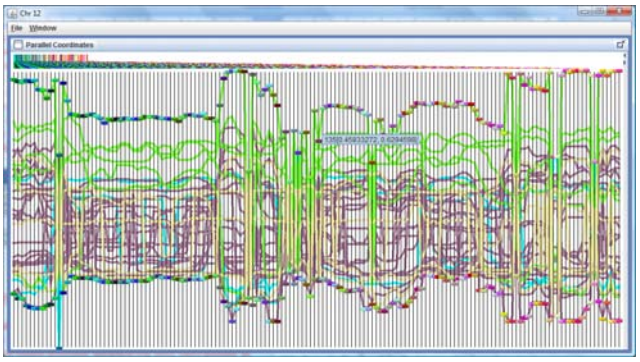
Once the data are represented, a CBR reasoning cycle is done. During the retrieve phase, the information about the catalogued genes and transcripts is recovered from the UCSC database. During the reuse phase, these genes are evaluated and valorized according to the hypothesis contrast started in section 3. Selecting the segments, the relevance of them can be observed. In figure 3, the information of the genes recovered from the database and considered as relevant can be seen.

Figure 3. Selection of segments and genes automatically



In order to facilitate the revision and learning phases to the expert, a different visualization of the data is provided. This view helps to verify the results obtained by the hypothesis contrast about the significance of the differences between pathologies. In figure 4, a representation in parallel coordinates is shown. Each line is associated to a patient and the color represents the pathology type. Each coordinate represents a segment. And the green lines are separated clearly from the rest, this means that the differences can be considered as important.

Figure 4. Selection of segments and genes automatically



The presented system facilitates the use of different sources of information for the analysis of the relevance in variations localized in chromosomic regions. The system is able to selects the genes that characterize pathologies automatically, using a CBR. This CBR allows the management of external sources of information for the generation of final results. The provided visualizations permit to validate the obtained results by an expert in an easier and faster way. If we compare the system with other proposals, it is possible to see that it facilitates the knowledge management of databases, thus it provides automatic methods for retrieving relevant information and providing reports.

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

References

Aha D., Kibler D., Albert, M.K. (1991) Instance-based learning algorithms. Machine Learning. vol. 6, 37-66.

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

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

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

Cohen, W.W. (1995) Fast effective rule induction. In Proceedings of the 12th International Conference on Machine Learning. Morgan Kaufmann. 115-123

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

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

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

Holmes, G., Hall, M., Prank, E. (2007) Generating Rule Sets from Model Trees. Advanced Topics in Artificial Intelligence. vol. 1747/1999, 1-12

Holte, R.C. (1993) Very simple classification rules perform well on most commonly used datasets. Machine Learning. vol. 11, 63-91

Kim, S.Y., Nam, S.W., Lee, S.H., Park, W.S., Yoo, N.J., Lee, J.Y., Chung, Y.J. (2005) ArrayCyGHt, a web application for analysis and visualization of array-CGH data. Bioinformatics. vol. 21(10),2554-2555

Kolodner J. (1993) Case-Based Reasoning. Morgan Kaufmann.

Lingjaerde, O.C., Baumbush, L.O., Liestol, K., Glad, I.K., Borresen-Dale, A.L. (2004) CGH-explorer, a program for analysis of array-CGH data. Bioinformatics, vol. 21(6), 821-822

Mantripragada, K.K., Buckley, P.G., Diaz de Stahl, T., Dumanski, J.P. (2004) Genomic microarrays in the spotlight. Trends Genetics. vol. 20 (2), 87-94

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. (2006) Array-CGHbase: an analysis platform for comparative genomic hybridization microarrays. BMC Bioinformatics. vol. 6 (124) 179-187

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

Quinlan, J.R. (1993) C4.5: Programs For Machine Learning. Morgan Kaufmann Publishers Inc.

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. (2006) VAMP: Visualization and analysis of array-CGH, transcriptome and other molecular profiles Bioinformatics. vol. 22 (17), 2066-2073

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

Takeishi, A. (2002) Knowledge Partitioning in the Interfirm Division of Labor: The Case of Automotive Product Development, Organization Science (13:3), 321-338.

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

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

Xia, X., McClelland, M., Wang, Y. (2005) WebArray, an online platform for microarray data analysis. BMC Bionformatics. vol. 6 (306), 1737-1745
Ylstra, B., Van den Ijssel, P., Carvalho, B. and Meijer, G. (2006) BAC to the future! or oligonucleotides: a perspective for microarray comparative genomic hybridization (array CGH). Nucleic Acids Research. vol. 34, 445–450

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

Hixson, P., Laritsky, E., Wang, X., Jiang, T., Cheung, S., Van Den Veyver, I., Cai, W. (2006) Comparison between BAC and oligo array platforms in detecting submicroscopic genomic rearrangements [abstract]. American Society of Human Genetics, Annual Meeting, 9-13, 2006, 239.

Hehir-Kwa, J.Y, Egmont-Petersen, M., Janssen, I.M., Smeets, D., van Kessel, A.G., Veltman, J.A. (2007) Genome-wide Copy Number Profiling on High-density Bacterial Artificial Chromosomes, Single-nucleotide Polymorphisms, and Oligonucleotide Microarrays: A Platform Comparison based on Statistical Power Analysis. DNA Research vol.14 (1) 1-11

Kenney, J. F. and Keeping, E. S. (1951) Mathematics of Statistics, Pt. 2, 2nd ed. Princeton, NJ: Van Nostrand,

Kruskal, W., and Wallis. W. (1952) Use of ranks in one-criterion variance analysis, Journalof American Statistics Association