Obtaining Relevant Genes by Analysis of Expression Arrays with a Multi-agent System

Alfonso González[a], Juan Ramos[a], Juan F. De Paz[a] and Juan M. Corchado[a]

{alfonsogb, juanrg, fcofds,corchado}@usal.es

[a] Biomedical Research Institute of Salamanca/BISITE Research Group, University of Salamanca, Edificio I+D+I, 37008 Salamanca, Spain

KEYWORD

Expression Arrays; Multi-agent System; CBR System; Pathway

ABSTRACT

Triple negative breast cancer (TNBC) is an aggressive form of breast cancer. Despite treatment with chemotherapy, relapses are frequent and response to these treatments is not the same in younger women as in older women. Therefore, the identification of genes that provoke this disease is required, as well as the identification of therapeutic targets.

There are currently different hybridization techniques, such as expression arrays, which measure the signal expression of both the genomic and transcriptomic levels of thousands of genes of a given sample. Probesets of Gene 1.0 ST GeneChip arrays provide the ultimate genome transcript coverage, providing a measurement of the expression level of the sample.

This paper proposes a multi-agent system to manage information of expres-sion arrays, with the goal of providing an intuitive system that is also extensible to analyze and interpret the results.

The roles of agent integrate different types of techniques, from statistical and data mining techniques that select a set of genes, to search techniques that find pathways in which such genes participate, and information extraction techniques that apply a CBR system to check if these genes are involved in the disease.

1. Introduction

There are several techniques that can be used to study genetic variation in patients, such as tissue microarrays, expression arrays (RNA) (Corchado et al, 2009; De Paz et al, 2008), genomic arrays (DNA) and arrays of microRNAs (miRNAs). Arrays used in the study of expression profiling are cDNA arrays and oligonucleotide chips. Moreover, different types of genomic arrays (DNA) are used, including BAC aCGH, oligo CGH, SNP CGH and aCGH (Comparative Genomic Hybridization) (Ylstra et al, 2006). CGH arrays (aCGH) can compare the DNA of a patient with control DNA and use this information to detect mutations (Pinkel et al, 2005; Mantripragada et al, 2004) based on the increase, loss or amplifications (Wang et al, 2005) in different regions of the chromosome. There are new exon arrays that provide accurate assessments of gene expression (Kapur et al, 2007). Information sources are varied but laboratory personnel usually follow fixed analysis processes that are distributed in sequences, which are in turn executed repeatedly in the search for genes that are considered to be relevant. Therefore, it is necessary to find a system that automates this process so that the work of the laboratory staff is simplified.

There have also been efforts to provide a solution to the main challenges associated with analyzing microarray data, which are: the high amount of data (coming from thousands of genes extracted from few samples); the high complexity of the data; the fact

that gene datasets in microarrays are often correlated (either directly or indirectly); and the fact that most gene selection and prediction models emphasize the capacity for effective classification instead of the function of an effective selection. The assumption is that statistical significance is equivalent to biological importance.

There are other investigations which focus their efforts in predicting genes that cause diseases. Thanh-Phuong Nguyne and Tu-bao Ho have developed a semi-supervised framework in order to find genes and detect possible connections among those that can lead to those diseases (Nguyen et al, 2012). They are based on feature extraction, preprocessing of data and integrate the following resources: Universal ProteinResource (UniProt) (UniProt, 2007) Gene Ontology (GO) (Gene Ontology, 2004), Pfam (Finn et al, 2008), InterDom (Ng et al, 2003), Reactome (Joshi-Tope et al, 2005) and to expression dataset (Dermitzakis, 2008). (Maglietta et al, 2007) propose a method from a similar point of view. The target is the selection of genes relevant to a pathology by analyzing the tissue expression profiles for two different phenotypic conditions. Statistical techniques are used and the presence of genes in similar studies is verified. Other studies use multiagent systems in order to analyze array data, including a system proposed by Juan F. De Paz et al. (De Paz et al, 2015), where a multiagent system analyses CGH arrays searching for gene gains or losses, which are then represented. This study is more oriented to obtain relevance areas and provide easy access to information but works only with CGH arrays. The present paper proposes a multi-agent system to analyze expression arrays. The main novelty is that the system can learn analysis flows (workflow) while the expression analysis is being performed, thus automating the analysis of expression. During the analysis, services are incorporated in order to carry out the analysis and extraction of information from database, through which most relevant genes are selected. Different data mining techniques and databases were used to analyze expression profiles and obtain relevant genes for two different phenotypic conditions. The system was applied to a real case study for the analysis of breast cancer with the aim of analyzing differences in this type of cancer with specific regard to the patient's age.

This article is organized as follows: section 2 describes the state of the art of expression arrays, Section 3 describes the proposal, and Section 4 presents the results and conclusions.

2. Gene Expression Arrays

Microarrays constitute a widely used tool that measure gene expression (Nuber, 2005). Moreover, this technology has attracted a special interest in cancer research (Knudsen, 2006). An expression arrays analysis makes it possible to study and compare transcriptomes of different samples. The value of gene expressions in these biochips is determined by the intensity of the hybridization of transcripts with a group of probes (Kapur et al, 2007).

With these qualities, expression arrays become a very useful tool that makes it possible to determine which genes have an altered expression, to compare expressions based on certain parameters, and to diagnose and distinguish subtypes of cancers with similar clinical manifestations, among other things. Different kinds of cancer genes share groups and altered pathways. Array analysis can investigate typical genes, as well as those that are not common to the vast majority of proliferative syndromes (Nuber, 2005), existing in more specific forms of the disease. This is one factor that makes arrays a useful diagnostic tool.

Beyond studying the expression of each gene and its degree of responsibility in an alteration, it is vital to understand the expression of these genes and the proteins they encode in the context of signaling pathways (Zhang et al, 2014).

To be able to perform a complete analysis, one of the roles of a multi-agent system is to search in different databases for the pathways taken by the specific genes that are being studied. One mutation in a particular gene can give rise to various effects, even in the same type of tissue (Vogelstein et al, 2004). Because of this, the function of the platform is interesting, specifically because the information obtained from the study of a single gene is not representative if, after it has been studied, its relationship to other elements that also influence the signaling pathway is not verified (Zhang et al, 2014).

The main function of this platform is to be able to select the relevant genes for the investigation. There comes a point during the screening process when there is no longer a sufficient number of elements to obtain pathways. It is precisely for this reason, from a

research point of view, that it is important to compare the genes obtained from the analysis that best explain the gene alteration (depending on the studied parameter) according to the altered pathways.

Among the most influential resulting gene expression analyses of patient samples in our case study, those that are also therapeutic targets are of particular interest to medical research. One of the great difficulties of the analysis of arrays is to obtain biologically valid conclusions from vast amounts of data (Lockhart et al, 2000). Consequently, one agent from the multi-agent system is responsible for conducting searches in databases known to contain therapeutic targets. This opens up the possibility of attacking these targets with drugs and conducting pharmacogenomic research after the analysis (Nuber, 2005). This makes it possible to check and directly study the influence of the pathway in the carcinogenic process, in addition to its clinical implications and the search for effective treatments.

3. Multi-agent System

In these studies, users have to work with a large volume of information, which involves the development of programs to improve data analysis systems and to automatically extract information through databases (Choon et al, 2014).

Our study uses expression arrays that determine the expression of genes to the probes used. This information is taken into account to observe differences that may occur in the same genes with regard to the age factor. Because large amounts of data are handled, it is necessary to develop a system aimed at simplifying the management and analysis of this information, and at automatically extracting information to determine the correlation of these genes in breast cancer.

Distributed analysis of expression data is performed by various personnel of the laboratory: from chip hybridization to the removal of variations and relevant information associated with the chips. This study shows a multiagent system specifically designed, with an abstract architecture for this virtual organizations (Argente et al, 2011), to analyze expression arrays. The functionality of the multi-agent system is divided into layers and roles to perform the analysis, which usually consists of several stages. The first stage is pre-processing, which performs the important task of the screening the data for the first time. The next stage performs an analysis of the expression probes, searching for differences with respect to the expression under normal conditions for that gene, or with respect to any specific factor. In the next stage data mining techniques are applied, allowing the data set studied to be further reduced. When looking for differences between groups of patients, it is important to confirm whether a cluster has been properly formed at the end

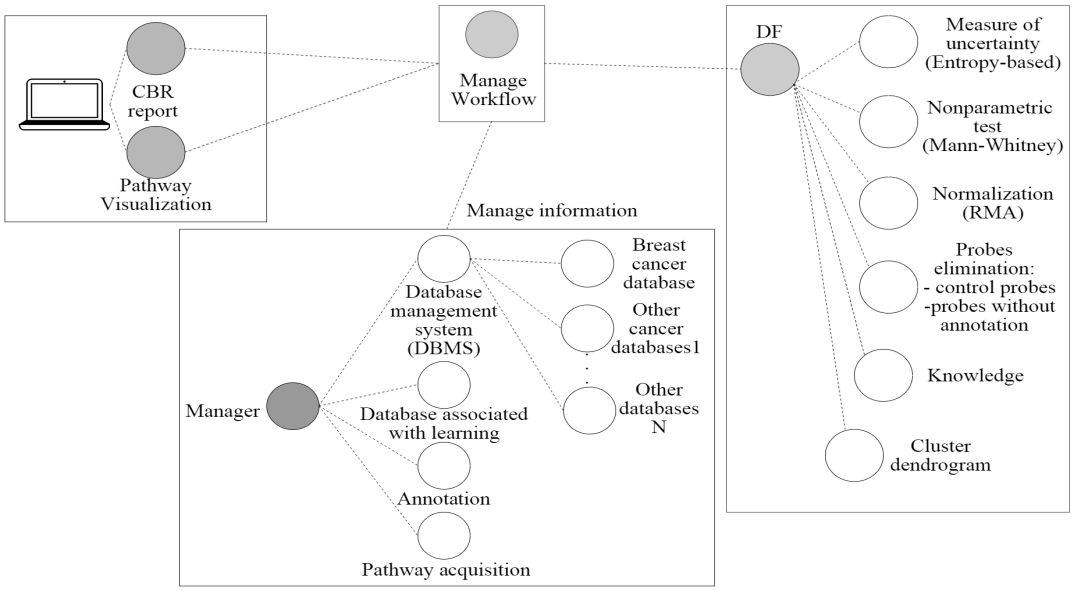

Fig. 1. Multi-agent System Arquitecture

of this stage, according to the case study. If a suitable result is not obtained, it will be necessary to review the previous step of extracting relevant genes through the data mining techniques. The final stage is initiated once the data set containing the different genes has been identified. This data set will be transferred to a database that checks the implication of these genes in the specific disease being studied to determine whether there is or is not a relationship.

JADE (Java Agent DEvelopment Framework) was adopted for the design and implementation of the proposed intelligent multi-agent and the architecture is composed of four layers: Analysis, Information Management, Visualization and Workflow. The architecture is shown in Fig.1.

The workflow layer includes an agent in charge of workflow in the other layers, and of establishing the correct order for the activity of each agent. Workflow analysis collects information about the settings and can repeat the sequences performed above for expression analysis. This aspect makes it possible to automate repetitive analysis tasks for laboratory personnel.

The analysis layer performs microarray analysis tasks as required by the process. This layer consists of several agents that are responsible for implementing the necessary processes and algorithms. An agent apply the Entropy-based filters, which is responsible for finding discrete attribute weights based on their correlation with continuous class attributes. Another agent applies the nonparametric Mann-Whitney statistical test to two independent samples. One more agent applies a normalization to adjust the signal, which may contain errors caused by technical and / or biological factors. The normalization technique applied is RMA (Robust multiarray Average), which adjusts information on each probeset to a comparable value. If any is over-expressed or under-expressed its genetic value is regulated with regard to the total. Another role performed by a system agent is the elimination of control probes and probes without notation. Finally, another agent produces a dendrogram that allows us to organize data into subcategories in which the case study is performed, this representation enables a clear representation of the relationship between data grouping.

The information management layer is responsible for confirming whether the genes obtained from the results of the analysis layer are related to the type of cancer that are the focus in the case study. This layer creates a database that collects the learned genes that are related to a cancer that is not contained in the databases used for the visit; this database is also consulted in the process of relating genes to the cancer in the case study. This layer also includes an Agent that collects the annotation and other information for each resulting gene and the agent that retrieves the pathways associated with these genes. The display layer manages the case based reasoning (CBR) system doing the four-step process. First the Multi-agent system retrieves cases relevant to solving the problem from the memory workflow. Once a relevant case is retrieved, that workflow is reused in the new problem with a previous adaptation as needed to fit the new problem. Later the revise step is performing to avoid undesired steps in the workflow for the next process: retaining the solution in a database.

The display layer also shows the pathway associated with the result as well as the list of the implicated genes. The display layer also manages the pathway visualization, showing the pathway associated with the gene list obtained. The gene list is sent using the KEGG Mapper (http://www.genome.jp/kegg/tool/map_pathway1.html). A pathway can be disrupted by a mutation on a single gene. The multi-agent system selects the pathways which present several changes in the pathology due to age factor. The selected pathways are shown.

4. Results

The case study was performed with 16 samples from patients with triple negative breast cancer provided by the Salamanca Cancer Institute. 8 samples corresponded to younger patients (less than 45 years) and the remaining 8 samples to elderly patients (over 68 years). Additional samples continue to be gathered in order to improve the accuracy of the results.

The technology used to analyze the microarray was Affymetrix, which is based on oligonucleotide chips. The specific chip used was the HuGene-1.0st-v1 chip, which contains 33,297 probes that identify about 23,000 sequenced genes.

The multi-agent system designed in this paper is applied to study gene expression arrays from the samples of these patients. The goal is to obtain the

genes that show differences between samples from younger patients and older patients in order to discover why older women respond better to the treatment.

The first step in the analysis of oligonucleotide chips is the process of discarding any control probes and probes without notation. Once these probes are discarded a denormalization process (Armstrong et al, 2004) is performed. This process discards the values that deviate from the normal value, and applies the statistic test (Mann-Whitney) to each of the independent samples used in the case study (younger women and older women). In that step a Benjamini and Hochberg FDR multiple test correction is performed and apply on the p-value calculated based on the two-samples Mann-Whitney.

In our case of study, we look for variations that may occur in the expression levels of genes for samples associated with younger women compared with those of the older women. Once applied, we discard all values exceeding a p-value greater than 0.01, i.e., we keep the probes that have a greater interest because of their expression level compared with older youth at statistical level.

In this process of data analysis, once the preprocessing, normalization and application of statistical tests and techniques of data mining are completed, a clustering algorithm is applied. This algorithm provides us with a dendrogram which allows us to check the degree of clustering of the probes with respect to the two samples (probes for younger and older women). Process results are shown in Fig.2.

The next stage is responsible for managing information received for the data obtained at the end of the analysis stage. These genes are contrasted with the corresponding type of cancer data base with which the data are associated (in this case study, the data base is for breast cancer). With this process we discard genes that are not implicated in a certain cancer, focusing on those that are.

The pertinent notation and associated pathways are obtained for this final gene set, allowing the access and use of this information.

The workflow layer agent learns the execution sequence of tasks, the order in which the agents interact with each other to execute the various processes and algorithms. With this initial learning, the following analysis of expression arrays are performed automatically, so the lab technicians do not have to perform the process manually, which avoids the risk of human error, loss of time, or a lower yield.

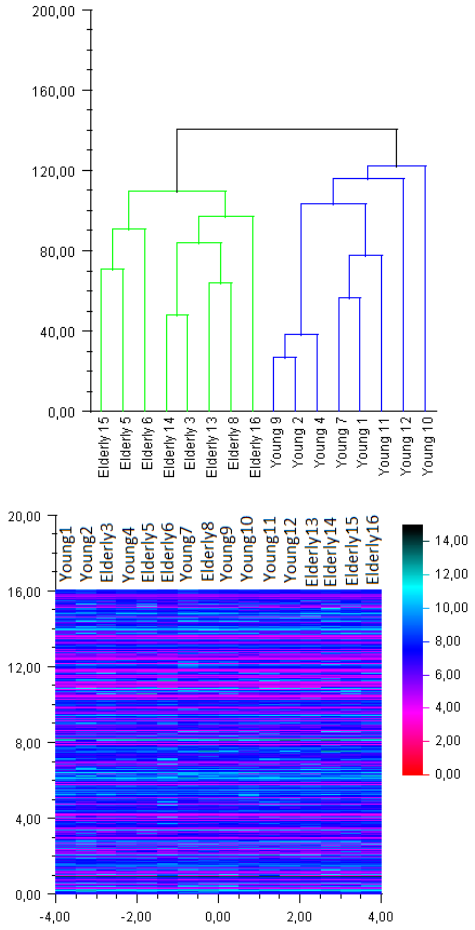

Fig. 2. Dendogram and Heatmap resulting

Once the genes and pathways are shown, an agent from the visualization layer performs a reasoning cycle (CBR). During the recovery step, information from the catalogued gene is obtained from the accurate database G2SBC (Genes-to-Systems Breast Cancer Database - http://www.itb.cnr.it/breastcancer/). While contrasting genes with the breast cancer database, these genes are evaluated according to the contrast hypothesis mentioned in part 3. In this way, if our system detects genes which are not in the databases and influence in the pathology according to the results, those genes are stored in an own database for future analysis

Table 1. shows the most important genes obtained in the case study for triple negative breast cancer. These genes are considered by the system as the most important in the existence of differences in response to treatment of younger women versus olderer women.

Table 1. Overexpressed genes resulting

Altered genes with close relationship to the age factor	**Altered genes with no-apparent relation to the age factor**
GABRP	CAPN6
SFRP1	SLC6A14
MID1	SCRG1
RARB	BCL11A
ACTG2	PTGS2
	BBOX1
	S100A7
	PRKAA2
	ACPP
	ALCAM
	RND3
	GGH
	PKP2

There was no pathway in which two or more selected genes were present. However, this is not surprising since the final number of genes kept is very small in order to provide a manageable quantity for the researcher. Although it is well known that expression varies with aging, there are few simultaneous gene relationships between age and breast cancer at the same time described previously in literature. Maybe cases like RARB are related to the age by way of methylation. Perhaps the clearest relationship previously described in literature (Elzi et al, 2012) is the case of SFRP1, which is an antagonist of the Wnt Signaling pathway overexpressed during senescence in response to DNA damage. Cellular senescence in young people acts as an antitumor mechanism.

5. Conclusions

The developed system enables using patient samples to know if there are differences in the expression level for the proposed gene sets, allowing the system to return the genes that produce differences in the samples with regard to the associated notation and pathways in which are implicated.

This case study looked at the differences in expressions that can occur in female patients younger than 50 years of age compared with those older than 50, since the latter group respond better to the treatments used.

This study is interesting because finding genes that behave differently can bring new information and the possibility of adjusting treatments for this type of cancer in younger patients.

The multi-agent system is developed in such a way that allows new agents to be inserted with new techniques or existing data to be modified for analysis. The system provides access to various databases so different cancer datasets can be introduced.

The system uses a CBR that handles all information obtained from the databases and allows the incorporation of new information that may be used in future analysis.

Acknowledgments. This work has been c has been supported by the Spanish Government through the project iHAS (grant TIN2012-36586-C01/C02/C03) and FEDER funds

6. References

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

Armstrong, N. J., & Van De Wiel, M. A.: Microarray data analysis: From hypotheses to conclusions using gene expression data. *Cellular Oncology, 26*(5-6), 279-290. (2004)

Choon, Y.W., Mohamad, M.S., Deris, S., Illias, R.M., Chong, C.K. and Chai. L.E.: A hybrid of bees algorithm and flux balance analysis with OptKnock as a platform for in silico optimization of microbial strains. Bioprocess and Biosystems Engineering. vol. 37 (3), 521- 532 (2014).

Corchado, J.M., De Paz, J.F., Rodríguez, S. and Bajo J.: Model of experts for decision support in the diagnosis of leukemia patients. Artificial Intelligence in Medicine. vol. 46 (3), pp. 179-200 (2009).

De Paz, J.F., Bajo, J., Vera, V. and Corchado J.M.: MicroCBR: A case-based reasoning architecture for the classification of microarray data. Applied Soft Computing. vol. 11(8) 4496-4507 (2011).

De Paz, J.F., Benito R., Bajo, J., Rodríguez-Vicente A., Abáigar M.: aCGH-MAS: Analysis of aCGH by Means of Multi-agent System. In press. Hindawi Publishing Corporation.

Dermitzakis, E.T.: From gene expression to diseaserisk. *Nature Genetics, 40 (5)*, pp. 492-493 (2008)

Elzi, D. J., Song, M., Hakala, K., Weintraub, S. T., & Shiio, Y.: Wnt Antagonist SFRP1 Functions as a Secreted Mediator of Senescence. *Molecular and Cellular Biology, 32*(21), 4388–4399. (2012)

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., Forslund, K., Eddy, S.R., Sonnhammer, E.L.L. and Bateman, A.: The Pfam protein families database. *Nucleic Acids Research*, *36*(Database issue), D281–D288. (2008)

Gene Ontology Consortium The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, *32*(Database issue), D258–D261. (2004)

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal,B., Gopinath, G.R., Wu G.R., Matthews, L., Lewis, S., Birney, E. and Stein, L.: Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, *33*(Database issue), D428–D432. doi:10.1093/nar/gki072 (2005)

Kapur, K., Xing, Y., Ouyang, Z., & Wong, W. H.: Exon arrays provide accurate assessments of gene expression. Genome Biology, 8(5), R82. (2007)

Knudsen, S. (2006). *Cancer Diagnostics with DNA Microarrays*. Wiley-Liss.

Lockhart, D.J. and Winzeler, E.A.: Genomics, gene expression and DNA arrays. Nature 405, pp. 827-836. (2000)

Maglietta, R., D'Addabbo, A., Piepoli, A., Perri, F., Liuni, S., Pesole, G. and Ancona, N.: Selection of relevant genes in cancer diagnosis based on their prediction accuracy. Artif Intell Med, 40(1):29-44. (2007)

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

Nguyen, T. P., & Ho, T. B.: Detecting disease genes based on semi-supervised learning and protein–protein interaction networks. *Artificial intelligence in medicine*, *54*(1), 63-71. (2012).

Ng, S.-K., Zhang, Z., Tan, S.-H., & Lin, K.: InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Research*, *31*(1), 251–254. (2003).

Nuber U.A. (2005) *DNA Microarrays.* New York, NY. Taylor & Francis group

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

The UniProt Consortium.: The Universal Protein Resource (UniProt). *Nucleic Acids Research*, *35*(Database issue), D193–D197. (2007)

Vogelstein, B., & Kinzler, K. W. Cancer genes and the pathways they control. *Nature Medicine*10, 789–799 (2004)

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

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

Zhang, J., Wu, L.-Y., Zhang, X.-S., & Zhang, S.: Discovery of co-occurring driver pathways in cancer. *BMC Bioinformatics*, *15*(1), 271. (2014)