Incorporating biological knowledge to microarray data classification through genomic data fusion

Daniel Glez-Peña[1]
Marta Pérez-Fernández[2]
Miguel Reboiro-Jato[3]
Florentino Fdez-Riverola[4]
Dpto. de Informática
ESEI: Escuela Superior de Ingeniería Informática
University of Vigo, Spain
[1]dgpena@uvigo.es
[2]mapfernandez@correo.ei.uvigo.es
[3]mrjato@uvigo.es
[4]riverola@uvigo.es

Marta Pérez
Dpto. de Matemáticas
ESEI: Escuela Superior de Ingeniería Informática
University of Vigo, Spain
martapr@uvigo.es

Fernando Díaz
Dpto. de Informática
EUI: Escuela Universitaria de Informática
University of Valladolid, Spain
fdiaz@infor.uva.es

Abstract - In this paper we propose the utilisation of an evolutionary approach for the task of classifying microarray data by using prior knowledge in the form of existing gene sets. The purpose of the work is to obtain an accurate classification model that uses a biologically relevant, and previously defined, gene set. The proposed algorithm will be integrated within geneCBR, a successful system able to perform data-mining over high-dimensional microarray data. Preliminary results show that the proposal is able to boost the biological relevance of candidate genes while maintaining the classification accuracy of the model.

Keywords: Microarray classification, genetic algorithm, biological relevance, gene sets.

1 Introduction and Motivation

Bioinformatics and medical informatics are two research fields that serve the needs of different but related communities. Both domains share the common goal to provide new algorithms, methods and technological solutions to biomedical research and contribute to the treatment and cure of diseases [1]. While bioinformatics has been traditionally focused on the intersection between computer science and biological research, medical informatics has been centered on the intersection between computer science and clinical medicine. In this context, recent studies have shown how biomedical informatics has emerged as a new area to describe the technology that brings both disciplines together to support genomic medicine [2].

From another perspective, it is commonly accepted that there are two parts to health sciences: (*i*) the study, research, and knowledge of health and (*ii*) the application of that knowledge to improve health, cure diseases, and understand how humans function. This configuration (theory elicitation and theory application) is analogous with the know-how managed by physicians that apply a mixture of objective knowledge (e.g. textbook information, medical databases, etc.) and subjective knowledge (e.g. experience, typical and exceptional recorded cases, etc.) [3].

Given the applied nature of medical informatics, bioinformatics and, more recently, biomedical informatics, the particularities regarding to the context retention, acquisition, representation, transferability and applicability of domain knowledge in the development of successful applications becomes critical. In these disciplines, the available knowledge spectrum spans from a complex reality to high level abstractions using classical concepts as data, meta-data, information, knowledge and meta-knowledge. While the tools developed for medical informatics reach a wide range of users (physicians, nurses, administrators, management, researches, etc.), bioinformatics applications are characterized by a much more homogeneous user group dominated by researches.

In the particular case of medical informatics and focused on fundamental aspects of decision-making, the work of [4] showed the convenience of using case-based reasoning (CBR) paradigm as a technical solution able to continuously process individual knowledge in order to advance towards personalized healthcare. As stated by the authors, advancing on individual knowledge processing provides an alternative solution to the problems that arise from the use of general knowledge. Nowadays, medical applications of case-based reasoning cover all aspects of health, being a wide and multidisciplinary research area expected to growth exponentially.

From the promising perspective of existing fruitful applications of CBR systems broadly applied to health sciences, we have developed geneCBR, a successful model that can perform cancer classification based on microarray data [5]. geneCBR employs a case-based reasoning model that incorporates a set of fuzzy prototypes for the retrieval of relevant genes, a growing

cell structure (GCS) network and a proportional weighted voting algorithm for the clustering of similar patients and the assignation of an initial class. The retrieval, reuse, revision and learning stages of our CBR system use these techniques to facilitate the CBR adaptation to the domain of biomedicine with microarray datasets.

Although numerical analysis of microarray data is quite consolidated, the true integration of numerical analysis and biological knowledge is still a long way off [7]. The inclusion of additional knowledge sources in the classification process can prevent the discovery of the obvious, complement a data-inferred hypothesis with references to already proposed relations, avoid overconfident predictions and allow us to systematically relate the analysis findings to present knowledge [8]. In this work we would like to enhance previous results obtained with geneCBR [9] in order to make interpretable predictions in concert with the incorporated knowledge. In this sense, our goal is to provide a way for the integration of imperfect biological knowledge with gene expression data and other high-throughput data sources.

The rest of the paper is structured as follows: Section 2 presents the proposed framework for integrating geneCBR with the manually-curated and fully traceable data derived from primary knowledge bases. Section 3 introduces the design and configuration of our preliminary experiments. Finally, Section 4 discusses the main issues as well as the future lines of our research work.

2 The proposed framework

In our previous geneCBR system, each class can be represented by a *fuzzy pattern* (FP) that can be constructed from the *fuzzy microarray descriptor* (FMD) associated with each one of the microarrays [10]. The FMD contains a value for each gene expression in terms of one of the linguistic labels 'Low', 'Medium' and 'High'. Therefore, each FP is a subset of genes based on the FMDs belonging to the same class, where the membership criterion of each gene to the fuzzy pattern of the class is frequency-based. In addition, a *discriminant fuzzy pattern* (DFP) of a set of FPs includes only those genes that can serve to differentiate it from the rest of the patterns. The algorithm used to compute the DFP given a collection of fuzzy patterns can be consulted in [11].

The set of discriminative genes identified by a DFP in our previous geneCBR system is the entry point in the framework proposed in the present work (top left in Figure 1). The main goal of the proposed framework is to define a systematic approach for incorporating biological knowledge in the form of gene sets coming from different databases (bottom part of Figure 1) without sacrificing gene-level classification accuracy of our previous geneCBR system. For this purpose we employ a genetic algorithm, a widely-used general-purpose search strategy for solving optimization problems [12].

The general process of the proposed model starts by establishing an initial population of individuals (chromosomes). Each individual represents a candidate subset of microarray genes, that is, a possible solution. Concretely, the initial individuals are randomly generated as subsets of the *discriminant fuzzy pattern* of the geneCBR system (right arrow labelled with number 1 in Figure 1). These chromosomes (having only differentially expressed genes, thus with high probability of presenting good performance in classification), will evolve using two genetic operators: crossover and mutation. These genetic operators alter the composition of chromosomes and create new solutions called offspring.

In the crossover operator, it generates offspring that inherits genes from both parents with a crossover probability P_c. In the mutation operator, it randomly selects genes for one of the most similar groups between the given chromosome and the biologically gene sets, and alter their status in the chromosome with a mutation probability P_m (right arrows labelled with number 2 in Figure 1).

The individuals belonging to the modified population need to be evaluated according to their utility for solving the optimization problem (right arrow labelled with number 3 in Figure 1). For this purpose, we define the fitness function as follows:

$$fitness = W_c \times CM + W_b \times BR \quad (1)$$

where W_c is the weight for the classification accuracy of a given classification model (CM), and W_b = (1 - W_c) represents the importance assigned to the biological relevance (BR) of this individual. Equation (1) modulates and summarizes the importance of data-based and biologically-based perspectives of a given chromosome (left arrow labelled with number 4 in Figure 1).

From the ranked set of individuals of a given offspring, the selection operator, inspired in natural selection fitter chromosomes survive and weaker ones die (left arrow labelled with number 5 in Figure 1). This step generates an improved population where the fitter chromosome has higher probabilities of being selected in the next generation (arrow labelled with number 6 in Figure 1). After several generations (executions of the whole cycle) the genetic algorithm is expected to converge to the best solution, the one with high classification accuracy and strong biological background.

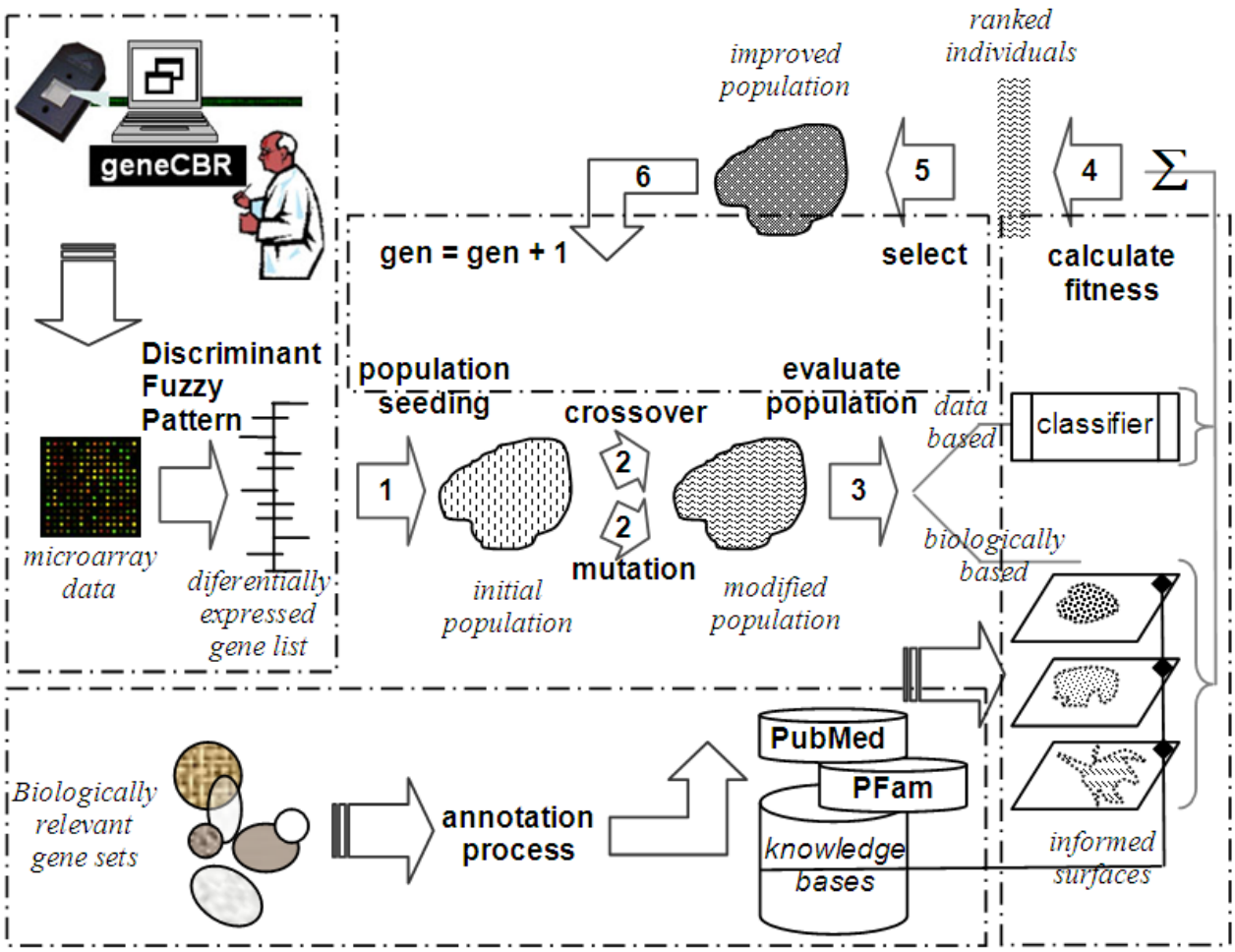

Figure 1 Flowchart of the proposed integrated model and its combinations with geneCBR system.

2.1 Representing prior knowledge as biologically relevant gene sets

In order to address the problem of acquiring, representing and applying domain knowledge about the analysed microarray data, we adopted the decision of representing gene groups whose constituents show subtle but coordinated expression changes. This idea was borrowed from recently developed gene set analysis methods that evaluate differential expression patterns of gene groups instead of those of individual genes [13].

Among the currently downloadable gene set analysis tools and databases, GSEA [14] is highly recommended for human gene expression dataset, with an enough number of samples (more than 10) [13]. In order to define a group of biologically relevant gene sets to integrate with geneCBR, we started from the CGP gene set collection (accessed on May 2, 2008) of the GSEA, which contains 1,186 gene sets that represent gene expression signatures of genetic and chemical perturbations. We have reduced this initial collection to 355 sets following the next steps: (*i*) choose the human related sets submitted by the Broad Institute, (*ii*) merge 9 duplicated groups and (*iii*) remove the groups showing an atypical high number of genes (greater than 120).

In our proposed model, biologically relevant gene sets are considered to directly affect the chromosomes (solutions of the genetic algorithm) through the utilisation of the mutation operator. The mutation process is executed for each chromosome on each cycle. Given a particular chromosome, a gene set presenting a higher level of similarity is randomly chosen. From this biologically relevant gene set, a number of genes (modulated by the mutation probability, P_m) are induced into the chromosome while others are deactivated in order to maintain the chromosome length stable.

Table 1 Informed surfaces for the probe sets of GeneChip HGU133A array.

	Terms	*GeneChip coverage*	*Annotations statistics*			
			Annotated probe sets	*Min*	*Max*	*Mean / Dev*
Pathway	199	27.98%	6234	1	32	2.6 / 9.49
Pfam	3217	83.79%	18670	1	10	1.68 / 1.12
GO	5585	87.88%	19583	1	60	7.98 / 24.17
OMIM	11253	73.04%	16276	1	13	1.28 / 0.68
Symbol	13108	96.24%	21446	1	1	1 / 0
PubMed	154840	95.45%	21269	1	1877	27.68 / 2913.27

2.2 Calculating the biological relevance through genomic data fusion

Today, multiple data sources are publicly available through Internet such as Gene Ontology (GO) annotations [15], the published literature [16], gene related databases (Entrez Gene, GeneCards), protein domain databases (Pfam, PROSITE, InterPro, SMART, Conserved Domain), protein interaction networks (STRING, IntAct), protein structure (PDB, SCOP, CATH), metabolic pathways (KEGG, BioCarta, Reactome, GenMAPP, PathwayInteractionDatabase), genomic sequence information (EMBL-Bank, DDBJ, Entrez Nucleotide), complete genomes (Entrez Genome, Ensembl, UCSC, TIGR), gene expression data (GEO, ArrayExpress) and a large etcetera [17].

With the growing number of complementary data and knowledge bases, the attention has shifted from the application of pure data-oriented methods to methods that aim to include additional knowledge in the data analysis process. This has translated into a need for sophisticated tools to mine, integrate and prioritize massive amounts of information [18].

In our approach, we provide the expert with control over the selection of biologically relevant gene sets and thereby taking advantage of the expertise of the user. Moreover, the developed model computes the similarity between the input gene sets and the genetic algorithm individuals by means of projecting them into *informed surfaces* and measuring the common overlapping area (right bottom in Figure 1). The objective of using informed surfaces is to provide the framework with the capacity of accessing multiple heterogeneous gene data sources (annotation, sequence, regulation, expression, etc.) in an integrated and homogeneous way.

In order to create the internal representation of the selected data sources, our model needs to know the structure of the microarray platform to be analysed.

Table 1 summarizes the data held by the selected informed surfaces for the case of Affymetrix GeneChip Human Genome U133A Array. For each database we give the number of identifiers, the percentage of array probes annotated with at least one identifier (coverage), and some annotation statistics of how the annotations are distributed among the probe sets. For the experiments carried out in this work we have selected de OMIM (*Online Mendelian Inheritance in Man*) informed surface in order to obtain conclusions about the validity of the existing knowledge to guide our informed search.

3 Experimental setup and preliminary results

In our previous experimentation [9], we worked with a database of bone marrow cases from 43 adult patients with Acute Myeloid Leukemia (AML) plus a group of six samples belonging to healthy persons for control purposes. Each case (microarray experiment) stores 22,283 ESTs corresponding to the expression level of thousands of genes. The data consisted of 1,025,018 scanned intensities belonging to the Affymetrix GeneChip Human Genome U133A Array.

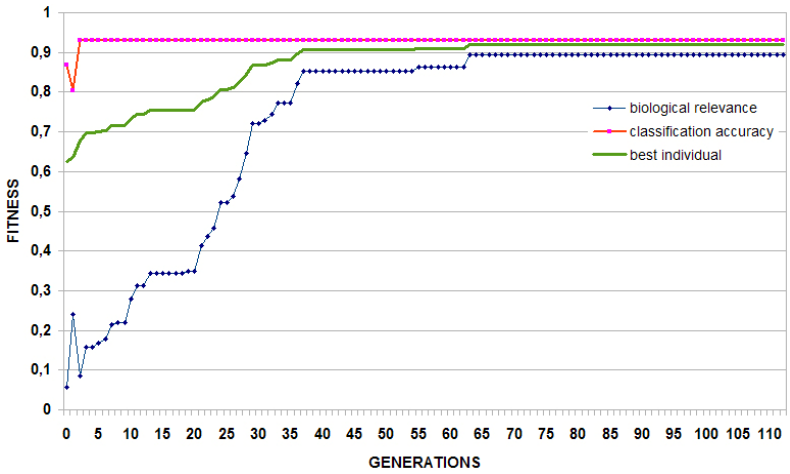

Figure 2 Evolution of the genetic algorithm using the OMIM informed surface.

The proposed model is applied over the same data set without taking into consideration the six samples belonging to healthy persons. For the present experiment, the initial population for the genetic algorithm was established to 20 individuals, where the dimensionality of each chromosome is 22,283 genes. The crossover probability, P_c, was set up to 0.5 and the mutation probability, P_m, was established to 0.0001. The strategy for selecting individuals between offspring was the tournament scheme. The maximum number of generations was 500 with a stop criterion of 50 generations without improvement.

The ratio between classification accuracy and biological relevance was set up to 0.7/0.3 and the biologically-based fitness function was configured to measure the similarity between gene sets taking into consideration the number of common OMIM (www.ncbi.nlm.nih.gov/omim/) entries. For the calculation of the affinity of each biologically relevant gene set with a given chromosome (individual) we used the Ochiai similarity coefficient [19].

In order to guarantee the quality of our experimental results, we have used a 10-fold stratified cross-validation for all the experiments [20]. The results reported in this study are those corresponding to the best execution of the 10-folds. From a general perspective about the global performance of the genetic search carried out, Figure 2 shows the evolution of the hybrid search. From Figure 2 it can be seen how from generation 65 the algorithm does not improve the relevance of the best individual (stopping after 50 iterations without improvement).

A graphical representation about the biological relevance of the relevant gene sets that conforms the best individual is shown in Figure 3.

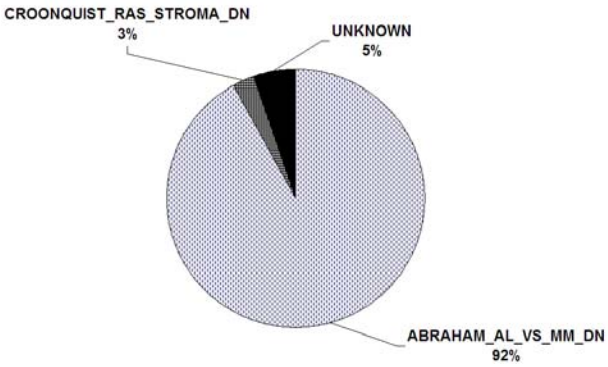

Figure 3 Graphical representation of the gene sets that explain the biological relevance of the best individual discovered by the genetic algorithm.

Figure 3 illustrates the effectiveness of the proposed method. From the 355 initial groups belonging to the CGP gene set collection of GSEA, only two groups were selected in order to explain the biologically relevance of the best individual. Moreover, the output of the model has shown other groups that are equivalent to the selected ones in terms of their biological relevance within the selected individual. Table 2 shows the equivalence between these two sets and those which are also similar to them.

Table 2 Other biologically relevant groups related with the best individual.

ABRAHAM_AL_VS_MM_DN	CROONQUIST_RAS_STROMA_DN
ABRAHAM_MM_VS_AL_UP	ASTON_DEPRESSION_UP
	CROONQUIST_IL6_STROMA_UP
	BRUNO_IL3_DN
	CORDERO_KRAS_KD_VS_CONTROL_UP

Another important output of the model is the list of genes that compose the best individual, as well as their comparison with the most accurate chromosome (from a classification perspective) belonging to the initial population (Table 3 shows this information). As it can be seen, the initial individual evolves from a starting classification accuracy of 86%, without any clear similar gene set, thus with a poor biological relevance (0.05), to a final individual having a very significant boost in its biological relevance (0.89) and an improved classification accuracy of 92%.

Table 3 Genes belonging to the best chromosome of the initial population and those belonging to the best individual discovered by the genetic algorithm.

Best initial individual [initial population] Fitness: *0.62444*		Best final individual [last generation] Fitness: *0.91922*			
classification accuracy *0.86747*	*biological relevance* *0.05736*	*Classification accuracy* *0.92949*	*Biological relevance* *0.89527*		
Probe set	**Gene Symbol**	**Probe Set**	**Gene Symbol**	**Probe Set**	**Gene Symbol**
201972_at	ATP6V1A	202763_at	CASP3	203460_s_at	PSEN1
208320_at	CABP1	211075_s_at	CD47	207782_s_at	PSEN1
207270_x_at	CD300C	213055_at	CD47	204261_s_at	PSEN2
219947_at	CLEC4A	213857_s_at	CD47	211711_s_at	PTEN
211839_s_at	CSF1	209953_s_at	CDC37	201783_s_at	RELA
207386_at	CYP7B1	202246_s_at	CDK4	209878_s_at	RELA
202942_at	ETFB	204247_s_at	CDK5	203084_at	TGFB1
206646_at	GLI1	203198_at	CDK9	203085_s_at	TGFB1
202957_at	HCLS1	219944_at	CLIP4	208864_s_at	TXN
221582_at	HIST3H2A	221582_at	HIST3H2A	201387_s_at	UCHL1
211799_x_at	HLA-C	201163_s_at	IGFBP7	215646_s_at	VCAN
213147_at	HOXA10	204863_s_at	IL6ST	205071_x_at	XRCC4
213926_s_at	HRB	204864_s_at	IL6ST	205072_s_at	XRCC4
201976_s_at	MYO10	211000_s_at	IL6ST	210812_at	XRCC4
212129_at	NIPA2	212195_at	IL6ST	210813_s_at	XRCC4
203857_s_at	PDIA5	212196_at	IL6ST	208643_s_at	XRCC5
203555_at	PTPN18				
202148_s_at	PYCR1				
204019_s_at	SH3YL1				
213247_at	SVEP1				

Finally we show in Table 4 the confusion matrix generated by the best individual proposed by the genetic algorithm. The performance was measured using a set of 50% never seen samples. The classification error was 14.29% and the kappa statistic achieved a value of 0.76.

Table 4 Confusion matrix for the best final individual proposed by the genetic algorithm.

	LAM_LAP	LAM_INV	LAM_M5	LAM_OTHER
LAM_LAP	4	0	0	0
LAM_INV	0	0	0	0
LAM_M5	0	0	3	0
LAM_OTHER	0	3	0	11

4 Conclusion and discussion

In this work we have presented an evolutionary model for the task of classifying microarray data by taking into consideration existing prior knowledge. The purpose of the work is twofold: (*i*) obtaining an accurate classification model and (*ii*) augmenting the biological relevance of the select genes and biomarkers.

From the initial experiments carried out it can be seen that the proposed approach obtains good individuals able to both (*i*) classifying the underlying data and (*ii*) capture prior knowledge though the utilisation of informed surfaces.

Despite the fact that promising results are obtained, an important effort is needed in investigating several aspects of the proposed framework. In this sense, we can highlight the importance of the mutation operator in charge of exploring the search space (biologically relevant groups identified by the user). New and reformulated forms of incorporating prior knowledge by means of the mutation operator are required and must be conveniently tested. Moreover, the integration of multiple heterogeneous data sources requires a more sounded study in order to generate a global ranking using order statistics. Finally, in order to correctly validate the global performance of the whole system, it is necessary to define several complementary strategies to create multiple cases of study.

To sum up, an important advantage of hybridizing gene set analysis with individual gene analysis is that the congruency between different datasets on the same biological question increases much more when compared at a gene set level than at an individual gene level. Investigation in this direction is demanding and promising.

Acknowledgements

This work was partially supported by the project *Development of biomedical applications* (09VIB10) from University of Vigo. D. Glez-Peña acknowledges Xunta de Galicia (Spain) for the program Ángeles Alvariño.

References

[1] F. Martin-Sanchez et al., *Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care*, Journal of Biomedical Informatics, Vol. 37, Issue 1, pp 30-42, February 2004.

[2] Dietrich Rebholz-Schuhman et al., *SYMBIOmatics: Synergies in Medical Informatics and Bioinformatics – exploring current scientific literature for emerging topics*, BMC Bioinformatics, Vol. 8, Suppl. 1, S18, March 2007.

[3] Rainer Schmindt et al., *Case-based reasoning for medical knowledge-based systems*, International Journal of Medical Informatics, Vol. 64, Issue 2-3, pp 355-367, November 2001.

[4] Stefan V. Pantazi, José F. Arocha and Jochen R. Moehr, *Case-based medical informatics*, BMC Medical Informatics and Decision Making, Vol. 4, No. 19, November 2004.

[5] Fernando Díaz, Florentino Fdez-Riverola, Juan M. Corchado, *gene-CBR: a Case-Based Reasoning Tool for Cancer Diagnosis using Microarray Datasets*, Computational Intelligence, Vol. 22, Issues 3-4, pp 254-268, November 2006.

[6] Daniel Glez-Peña, Fernando Díaz, Jose M. Hernández; Juan M. Corchado, Florentino Fdez-Riverola, *geneCBR: a translational tool for multiple-microarray analysis and integrative information retrieval for aiding diagnosis in cancer research*, BMC Bioinformatics, 10:187, June 2009.

[7] Francesca Cordero, Marco Botta and Raffaele A. Calogero, *Microarray data analysis and mining approaches*, Briefings in Functional Genomics and Proteomics, Vol. 6, Issue 4, pp 265-281, January 2008.

[8] Riccardo Bellazzi and Blaž Zupan, *Methodological Review: Towards knowledge-based gene expression data mining*, Journal of Biomedical Informatics, Vol. 40, Issue 6, pp 787-802, December 2007.

[9] Fernando Díaz et al., *Applying GCS networks to fuzzy discretized microarray data for tumour diagnosis*, Proceedings of the 7th International Conference on Intelligent Data Engineering and Automated Learning 2006, Burgos Spain.

[10] Florentino Fdez-Riverola et al., *Improving gene selection in microarray data analysis using fuzzy patterns inside a CBR system*, Proceedings of the 6th International Conference on Case-Based Reasoning 2005, Chicago - Illinois, August 2005.

[11] Fernando Díaz et al., *Using fuzzy patterns for gene selection and data reduction on microarray data*, Proceedings of the 7th International Conference on Intelligent Data Engineering and Automated Learning 2006, Burgos – Spain, September 2006.

[12] Mitsuo Gen and Runwei Cheng, *Genetic algorithms and engineering design*, John Willey & Sons Inc., New York, 1997

[13] Dougu Nam and Seon-Young Kim, *Gene-set approach for expression pattern analysis*, Briefings in Bioinformatics, Vol. 9, Issue 3, pp 189-197, January 2008.

[14] Aravind Subramanian et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles*, Proceedings of the National Academy of Sciences of the United States of America, Vol. 102, No. 43, pp 15545-15550, October 2005.

[15] Evelyn Camon et al., *The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT*, TrEMBL, and InterPro, Genome Research, Vol. 13, Issue 4, pp 662-672, March 2003.

[16] Pubmed, http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed, accessed December 2009.

[17] Alex Bateman, *Database Issue: Editorial*, Nucleic Acid Research, Vol. 36, Issue D1, January 2008.

[18] Stein Aerts et al., *Gene prioritization through genomic data fusion*, Nature Biotechnology, Vol. 24, Issue 5, pp 537-544, May 2006.

[19] Jair Moura Duarte, João Bosco dos Santos and Leonardo Cunha Melo, *Comparison of similarity coefficients based on RAPD markers in the common bean*, Genetics and Molecular Biology, Vol. 22, No. 3, pp 427-432, 1999.

[20] Ron Kohavi, *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*, Proceedings of the 14th International Joint Conference on Artificial Intelligence 1995, Canada August 20-25, Vol. 14, No. 2, pp 1137-1145.