Proceedings of the 2007 IEEE Symposium on Computational
Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

GoFuzzKegg: Mapping Genes to KEGG Pathways Using an Ontological Fuzzy Rule System

Mihail Popescu, *Member, IEEE*, Dong Xu, *Member, IEEE* and Erik Taylor

Abstract — **In this paper we present a method for finding the main pathways represented in a set of genes (say obtained from a microarray experiment). The method is based on a fuzzy mapping between genes represented as sets of Gene Ontology terms and KEGG pathways using a new type of fuzzy rule system called ontological fuzzy rule system (OFRS). As opposed to a crisp mapping, the fuzzy mapping produces a non-zero value even if the gene name is not explicitly listed in a given KEGG pathway. An OFRS is a fuzzy rule system in which the rule memberships are obtained using similarity measures between objects computed based on the Gene Ontology (GO) annotations. To test our approach, we randomly selected without replacement 10 sets of Arabidopsis thaliana genes from KEGG (each set had 15 genes from 3 different pathways) and tried to predict the pathways they were selected from. Our method was able to find, 90% of the right pathways with a 65% false alarm rate at a p-value of 0.01. The high false alarm rate is due in part to the experimental setting. In a pilot dataset of 526 Arabidopsis thaliana genes we identified 8 clusters which proved to be linked to important pathways such as ATP synthesis and transcription factor.**

I. INTRODUCTION

FUZZY computing differs from conventional (hard) computing in that, unlike hard computing, it is tolerant to uncertainty and partial truth. The model for fuzzy computing is the human mind. For example, we all know what "small" is but we probably disagree on the exact number of inches it represents.

Use of fuzzy techniques such as fuzzy clustering [1], fuzzy neural networks [2], fuzzy rule systems [3], fuzzy relations [4] and fuzzy rule systems [5] has been previously reported in bioinformatics. Fuzzy rule systems were also employed in other domains such as control theory and decision making [6].

The association between fuzzy techniques and ontologies has also been previously reported. Andreasen [7] introduced the concept of an ontological query in the context of fuzzy databases. Fuzzy ontologies were used by Tho [8] and Lee [9] in text mining and text summarization applications.

This work has supported in part by NLM grant 2-T15-LM07089-14.

M. Popescu is with the University of Missouri, Columbia, MO, 65211, USA (phone: 573-882-1266; e-mail: popescum@missouri.edu).

D. Xu is with the University of Missouri, Columbia, MO, 65211, USA (e-mail: xud@missouri.edu).

E. Tylor is with the University of Texas, Houston, TX, USA .

In this paper we use the fuzziness intrinsic to a crisp ontology (as in [7]) rather than building a fuzzy ontology per se (as in [8]).

Among bioinformatics ontologies (such as EcoCyc, Tambis, Gene Ontology (GO), MBO, KEGG) Gene Ontology seems to be used the most in applications. Among the multitude of applications where a GO based similarity between gene products has been reported, we mention [10]-[13]. In this work we will also use the GO for computing the similarity between gene products. In addition, our fuzzy logic approach to using Gene Ontology should be considered in the context of the knowledge driven approaches to GO such as the ones presented in [15] and [16].

Many applications, such as [16]-[18], provide a mapping of a set of genes to a set of pathways. By clustering the genes based on their pathway fuzzy memberships, we are able to compute not only the pathways associated with a group of genes, but also which gene subgroup is associated to which pathway.

II. METHODS

A. Ontological Similarity as a Fuzzy Membership

Let us assume that each gene product G is described by a set of GO terms as $G=\{T_1,\ldots,T_n\}$. The key concept in the fuzzy mapping of GO to KEGG is the gene product similarity. This similarity is computed using an ontology and ontology related algorithms. As observed by Andreasen [7], one can interpret the similarity between two ontology terms as a fuzzy membership of one term in the concept denoted by the other term. Consequently, the similarity between two genes may be used as a fuzzy membership. The fuzzy membership can be further employed to fire fuzzy rules. However, the above interpretation must be made with care.

First, while the similarity measure in an ontology is usually symmetrical, that is $s(a,b)=s(b,a)$, the fuzzy memberships might not be. The reason resides [7] in the difference between generalization ("serine-threonine kinases are kinases") and specialization ("kinases are serine-threonine kinases"): while the first holds totally (high membership value) the latter holds only partially ("medium" membership value).

Second, the high ontological relatedness between siblings of a term (say, tyrosine kinase and serine/threonine kinase)

might not translate in a high level of similarity between their functions. For example, only tyrosine kinases are growth factors and not serine/threonine kinases. We mention here that approaches to computing ontological similarity based on information content [20] are inadequate for our purpose since they are totally ignore the above issues. While we leave this topic for further research, in this paper we chose to calculate the term similarity (hence the fuzzy memberships) using the approach presented by Andreasen [7] that deals with the above problems.

In this approach, the similarity between two GO terms, T_1 and T_2, is computed as:

$$s_{12} = \max_{\{P_i\}} \prod_{j \in P_i} w_{ij}\,, \tag{1}$$

where $\{P_i\}$ is the set of all possible paths connecting T_1 and T_2 in GO, and w_{ij} is the weight assigned to the arc j from path P_i. (see figure 1).

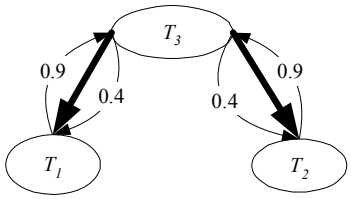

Figure 1. Example of path-based computation of the similarity (membership) between two GO terms. Note that the thin arcs represent the weight assignment process while the thick arcs represent the ontological relation "is-a"(GO is a directed acyclic graph). Here, the similarity between T_1 and T_2 is 0.9*0.4=0.36.

We mention that s_{12} is not in fact a similarity relation because it lacks the symmetry. A more complex formula that defines a similarity relation based on the same approach can be found in [7]. However, we believe that the lack of symmetry is, aside of simplicity, more suitable for our purpose in this work.

We consider only two types of weights here: specialization weights (downward from the ancestor node to the descendent node) with a value of 0.4 and generalization weights (upward from the descendent to the ancestor) with a value of 0.9. This assignment reflects the fact that "all T_1 are T_3" but only "some T_3 are T_1". For example, in figure 1, the similarity between and T_1 and T_2 is 0.9*0.4=0.36.

B. *Ontological fuzzy rule system (OFRS)*

A typical Mamdani fuzzy rule system (FRS) with one input variable in the antecedent (left side) and one output in the conclusion (right side) has the form [6]:

$$\begin{aligned} &\text{Rule 1: IF } x \text{ is } G_1 \text{ THEN } y \text{ is } P_1 \\ &\dots \\ &\text{Rule m: If } x \text{ is } G_m \text{ THEN } y \text{ is } P_m \end{aligned} \tag{2}$$

where G_i and P_i, i=1…m, are fuzzy sets, x is the input variable and y is the output variable. Fuzzy sets G_i are possible "values" for the variable x while P_i are possible "values" for y. The fuzzy sets are usually represented using membership functions. For example, in figure 2 we show three membership functions for the fuzzy variable "stature" called "short", "average" and "tall". To fire a rule, one has to compute the membership value (called rule activation) of a given value of the input variable x, $x_0 \in R$. For instance, in figure 2, for x_0=1.8 we get the membership values of 0 in "short", 0.2 in "average" and 0.8 in "tall".

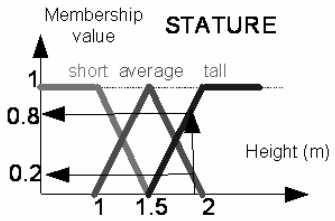

Figure 2. Memberships for fuzzy variable "STATURE": "short", "average", "tall".

The rule activations w_i are then used as weight factors in computing the output of the fuzzy rule system as:

$$y_0 = \underset{i=1,m}{Agg}(w_i P_i)\,, \tag{3}$$

where "Agg" is some aggregation operator (usually weighted average. The entire inference process is shown schematically in figure 3.

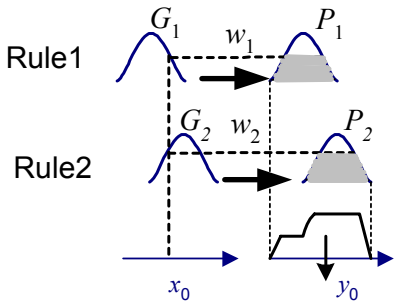

Figure 3. A typical Mamdani FRS with one input variable and two rules. The activation of Rule1, w_1, is given by the membership of x_0, $G_1(x_0)$. Similarly, the activation of Rule2, w_2, is given by $G_2(x_0)$. The system output, y_0, is calculated by aggregating the activated portion of the output membership of each rule (grayed area) using formula (3).

We note that if the above fuzzy rules (2) have the same output fuzzy set (say P_1) then they can be equivalently written as:

IF x is G_1 OR …OR x is G_n THEN P_1 (4)

In this case the rule activation w_{rule} can be computed using an OR operator [6] as:

$$w_{rule} = \underset{i=1,n}{OR}(w_i)\,, \tag{5}$$

In this paper we use as OR the "maximum" operator.

An ontological fuzzy rule, OFRS, is a Mamdani FRS (2) where the membership are computed using ontological similarities. It is possible that each variable has a different ontology associated. In our case, the input variables are genes annotated with terms from the Gene Ontology (GO) while the output variable is a KEGG pathway. The OFRS

for the GO to KEGG mapping can be represented as:

IF *gene* is $GENE_1$ OR … OR *gene* is $GENE_n$ (6)
THEN *pathway* is $PATH_1$

where $GENE_1$ to $GENE_n$ are identified by KEGG as being present in pathway $PATH_1$. We note that the above OFRS has as input variable one gene. The output of the OFRS is the membership of that gene in a pathway $PATH_1$. However, it is possible to build an OFRS with an input x that consists of a group of genes. In this case, the G_i's from formula (2) consist in all the genes from a pathway. The challenge of this approach is that we do not know *apriori* which group of genes is associated with which pathway.

Since (6) is an ontological FRS, the membership calculation in the rule antecedent are performed using the GO similarities between genes. The process is metaphorically represented in figure 4 (relate to figure 3). The membership functions from figure 3 were replaced by ontological trees in figure 4. Similarly, the output is computed (formula 3) by aggregating the KEGG pathways using KEGG similarity and the rule activations as weights.

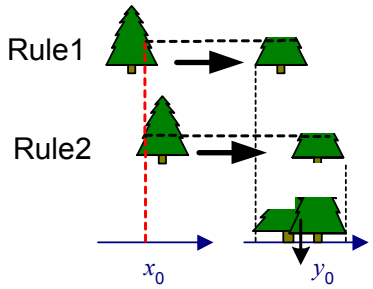

Figure 4. Metaphorical representation of a OFRS. The membership functions are replaced by similarity measure in ontological trees.

In this paper we do not use the KEGG pathway similarity in computing the OFRS output in formula (3). We leave this topic for further research. Instead we set as the output the pathway that has the highest activation. We mention that by taking into account the similarity between the KEGG pathways, the result of our testing (90% correct pathway prediction) may be in fact better. In some cases, we predict a pathway that is "very similar" to the real one but we still count it as a false prediction.

The input of the OFRS is the set of GO annotations for a given gene and the output is the gene membership in a given pathway. The degree of membership, y_0, of a gene in a pathway is calculated by firing the above rule (6) using the similarity of the gene with the genes in the pathway, $s(gene, G_i)$, in a similar fashion to (5) as:

$$y_0(gene, PATH_1) = \underset{i=1,n}{OR}\, s(gene, G_i) \cdot \quad (7)$$

where n is the number of genes in KEGG pathway $PATH_1$. Obviously, the case when the gene is explicitly mentioned in a given pathway is not very interesting (the rule outputs a membership equal to 1). However, if the gene is not explicitly mentioned, the membership y_0 will reflect the degree of functional homology to the genes from a given pathway. In (7) one could replace the OR operator with an OWA operator (21) by averaging the highest 3 scoring similarities in a pathway.

Since we have a rule for each pathway, a given gene will be described by a feature vector with length m (the number of pathways). For example, in KEGG there are m=181 pathways for *H. sapiens* and m=115 for *A. thaliana*. In fact, our fuzzy rule base consists in the KEGG pathway database for the given organism.

C. GoFuzzKegg algorithm

The input to our algorithm is a set of GO annotated genes $\{G_i\}_{i=1,N}$. We mention that the approach can be used even for un-annotated genes by first employing one of the automatic annotation methods [22]-[24]. The pathway prediction algorithm has the following steps:

Step 1. Compute the activation y_{ij} of each gene G_i, $i \in [1,N]$ in pathway j, $j \in [1,m]$ using (7). For our *A. thaliana* example m=115. For our test dataset, N=15 and for the pilot dataset N=526 (see next section). As a result each gene i is described by a pathway activation (feature) vector $Y_i=\{y_{i,1}, ..., y_{i,115}\}$ of size 115.

Step 2. Compute the gene similarity matrix, $S=\{s_{ij}\}_{i,j \in [1,N]}$, as:

$$s_{ij} = \frac{Y_i^T \bullet Y_j^T}{\sqrt{|\, Y_i^T \,||\, Y_j^T \,|}}, \quad (8)$$

where Y^T denotes that the vector Y was thresholded with a threshold T (that is, if $y_{i,k}<T$, then $y_{i,k}=0$} that will be determined later. The thresholding operation was performed in order to remove the noise (pathways with residual activation).

Step 3. Use a cluster validity measure to assess the most likely number C of pathways (clusters) present in the dataset. One could use the VAT algorithm [25] to visually assess the number of clusters in S. Other cluster validity measures, such as Dunn index or partition coefficient [26], may be used in conjunction with using the fuzzy C-means [27] algorithm to cluster the genes represented by the feature vectors $\{S_i\}_{i=1,N}$ into C clusters, where $S_i=\{s_{i1}, \ldots, s_{iN}\}$. We found that it is more reliable to cluster the matrix S using fuzzy C-means rather than the feature vectors $\{Y_i\}$ directly.

Step 4. Assume $I_k = \{i_1, ..., i_{|I_k|}\}$ is the set of indices from cluster $k \in [1,C]$, where $\sum_{k=1}^{C} |\, I_k \,| = N$ and $|I|$ denotes the cardinality of I.

To determine the pathways most likely associated to cluster k we sum all the genes in the cluster k as:

$$Sum_k = \sum_{i \in I_k} Y_i, \quad (10)$$

and then we compute the most likely pathway as:

$$p_k = \arg\max_{j=1,m}\{Sum_{kj}\}, \quad (11)$$

that is, we find the pathway with max sum activation for the genes in cluster *k*. To produce more than one candidate for cluster *k*, say *Q*, we can consider the pathway that has the second highest sum activation in the cluster, and so on. Other summarization strategies are possible here [28] but we found this simple method to work well.

Step 5: For the testing case, evaluate prediction error. The detection rate (*DR*, sensitivity) is computed as:

$$DR = \frac{no\_pathways\_correct\_predicted}{total\_no\_correct\_pathways}, \quad (12)$$

The false prediction rate (FPR) is computed as:

$$FPR = \frac{no\_pathways\_erroneusly\_predicted}{total\_no\_pathways\_predicted}. \quad (13)$$

For example if the KEGG id's of the correct pathway are {10, 940, 3050}, and our prediction is {10,940, 3030, 4070}, then *DR*=0.66 and *FPR*=0.5. We reiterate the fact that we ignore that the pathways 3050 and 3030 are strongly related making our *DR* estimate more conservative.

To compute the p-value of our *DR* prediction, we assign the membership of the *N* genes in *C* clusters randomly and compute again *DR**. We perform the random assignment 1000 times, resulting in a set of 1000 random detection rates, $\{DR_j^*\}_{j=1,1000}$. Then, the p-value is calculated as:

$$p-value = \frac{\{no\_of\_DR_j^*, DR_j^* > DR\}}{1000}, \quad (14)$$

that is, the number of the random detection rates higher than our *DR* (obtained by clustering Y_i's) divided by 1000. The p-value is somewhat a measure of the reliability of our classifier: if the p-value is low (say, <0.05), a low detection rate might denote a hard gene set to predict and not a bad prediction method.

III. Datasets

As mentioned above, in this paper we used the KEGG pathway database for *A. thaliana* as our fuzzy rule database.

The test dataset consisted in 10 sets of 15 genes each, randomly selected (without replacement) from KEGG pathways (3 pathways for each set) that have more than 50 genes. The reason for this condition was that we tried to minimize the impact on the whole pathway at the extraction of 5 genes from it. We found 23 such pathways.

We mention that our approach does not need training data. This fact is one of the advantages of using fuzzy rule systems. Usually, the fuzzy rules are set up by domain experts. In our case, the memberships of genes in pathways (the rule base) were determined by biologists and stored in the KEGG database.

The pilot dataset that we used for further testing of our method consisted in 526 *A. thaliana* genes selected in a microarray experiment.

IV. Results

A. Finding the optimum pathway feature threshold T (formula 8)

To find the optimum value of T we computed the mean detection rate, DR, for all 10 sets of genes from the test set at different thresholds. To exclude the influence of the number of clusters on the detection rate we used a constant number of clusters N=3 (the number of pathways we selected in each set). The results were summarized in table 1.

TABLE I
DETECTION RATE, DR, FOR PATHWAYS IN THE TEST SET FOR DIFFERENT FEATURE THRESHOLDS, T.

Threshold (T)	0.1	0.3	**0.5**	0.7	0.9
Mean DR	0.43	0.47	**0.6**	0.56	0.47

From table 1 we see that the maximum detection rate was obtained for T=0.5. We will use this threshold for all our subsequent experiments.

B. Pathway prediction using one candidate pathway for each cluster

The results obtained on the 10 gene sets randomly selected from KEGG are shown in table 2. The prediction was obtained using one candidate pathway (the one that had the maximum activation sum) per cluster and using a feature threshold T=0.5.

TABLE II
PATHWAY PREDICTION RESULTS FOR THE TEST SET USING ONE CANDIDATE PATHWAY PER CLUSTER.

Set #	#pathways Predicted, C,(out of 3)	DR	FPR	#genes in right pathway (out of 15)
1	3	0.67	0.33	9
2	5	0.67	0.60	4
3	5	1.00	0.40	7
4	3	0.67	0.33	10
5	3	0.67	0.33	5
6	3	1.00	0	13
7	3	0.33	0.67	3
8	4	0.33	0.75	2
9	3	0.67	0.33	9
10	4	0.67	0.50	5
Avg		**0.66**	**0.43**	**6.7**

Note: The p-value was <0.01 for all 10 cases.

As we can see from table II, over-clustering (like in the sets number 2, 3, 8 and 10) leads to an increase in false predictions. Sometimes, clusters may be merged if they predict the same pathway. However, we leave pruning strategies for further research.

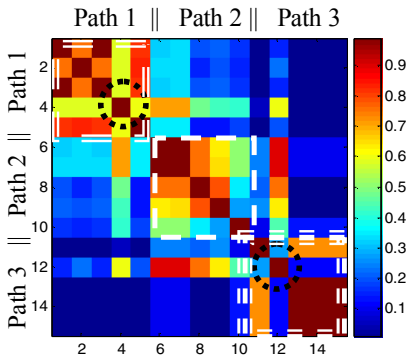

Figure 5. Similarity matrix between the 15 genes selected in case 6 from table 2. Genes 4 and 12 (circled) will be classified (erroneously) by fuzzy C-means together with the genes from pathway 2, {6,7,8,9,10}, instead of pathways 1 and 3, respectively.

We mention that by predicting the right pathways (like for set number 6) does not necessary mean that we assigned all the genes to the correct pathways in the process. For example in set number 6, we assigned only 13 out of 15 genes (87%) to the correct pathways. In figure 5 we show the gene similarity matrix computed using (1) and the pathway features for the set number 6. It can be seen that genes number 4 and 12 (circled) exhibit more similarity to the genes from pathway 2 (gene index 6 to 10- single line) than to their own pathways (gene index 1 to 5 –double line, and gene index 11 to 15-triple line, respectively).

In average, we predicted about 45% of the genes in the right pathway.

C. *Pathway prediction using multiple candidate pathways per cluster, Q*

We expect that increasing the number of candidate pathways per cluster, Q, will increase both the detection rate, DR, and the false prediction rate, FPR. This is exactly what we see in table 3, where the mean DR and mean FPR were recorded for a number of candidates Q∈[1,4].

TABLE III
THE VARIATION OF THE PATHWAY DETECTION RATE, *DR*, AND PATHWAY FALSE PREDICTION RATE, *FPR*, WITH THE NUMBER OF CANDIDATE PATHWAYS PER CLUSTER, *Q*.

Q	1	2	3	4
DR	0.66	0.8	0.84	0.9
FPR	0.43	0.58	0.60	0.64
p-val	0.001	0.01	0.01	0.02

D. *Results on the 526 A. Thaliana gene set*

Out of 526 genes we found only 438 to be annotated using a GO term. Since we did not use any automated annotation software in this work, we removed the 88 un-annotated genes from the experiment. To determine the most probable number of clusters we used the partition coefficient [26] that resulted in *C*=8 group of genes. Three representative pathways (*Q*=3) were then computed for each cluster (see table 4).

TABLE IV
THERE CANDIDATE PATHWAYS FOR EACH OF THE 8 CLUSTERS FOUND IN THE A. THALIANA DATASET.

Clst.	Size	Path 1 id	Path 2 id	Path 3 id
1	7	190	193	195
2	34	4130	53	3022
3	127	4710	230	280
4	56	940	600	903
5	25	3030	3022	3010
6	59	10	100	130
7	61	632	600	4130
8	69	280	290	770

We see that most of the clusters are coherent, that is, the pathway candidates for a cluster are very similar. For example, cluster 1 has 7 genes and the candidate pathways are: oxidative phosphorilation (190), ATP synthesis (193) and photosynthesys (195) (which are obviously related since 193 is included in 190 and 195 and 190 are both related to the energy metabolism). Similarly, cluster 5 has 25 genes and the candidates pathways are: DNA polymerase (3030), transcription factor (3022) and ribosome (3010) which are all involved in the DNA replication process. Finally, cluster 8 has 69 genes involved in "valine, leucine and isoleucine" degradation (280) and biosynthesis (290).

The similarity matrix for the 438 genes is shown below. One can clearly see the 8 clusters. Looking at figure 6 one could have further ideas of merging clusters. For example, the genes in cluster 4 and cluster 7 seem to be similar. Table 4 further confirms the observation since they share the second candidate: sphingolipid methabolism (600).

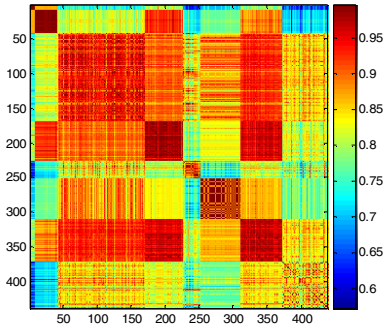

Figure 6. The pathway similarity matrix between the 438 A. thaliana genes. The matrix has been rearranged using the clusters obtained by applying fuzzy C-means on the initial similarity matrix.

V. CONCLUSIONS

In this paper we present a method for finding the pathways represented by a group of genes. The method is based on the fuzzy mapping of the Gene Ontology to KEGG pathways implemented as an ontological fuzzy rule system.

We tested our method on 10 test sets of 15 genes each extracted randomly from KEGG. The algorithm was able to predict 90% of the pathways with a 65% false prediction

rate. The high false alarm rate can be in part explained by the nature of the test that extracted without replacement 5 genes from each pathway. Having only few genes available for each pathway, the process of merely retrieving an extra gene results in 20% false alarm rate. However, extracting more genes from a KEGG pathway would lead to changing the nature of the pathway. To better test our algorithm we plan to use a microarray gene set fully investigated by biologists, where the genes-to-pathway mapping is known.

VI. ACKNOWLEDGEMENT

The authors wish to thank to Trupti Joshi for providing the *A. Thaliana* data.

REFERENCES

[1] Wang J, Bo TH, Jonassen I, Myklebost O, Hovig E. "Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data", BMC Bioinformatics, 4(1):60, Dec 2003.

[2] Ando T, Suguro M, Hanai T, Kobayashi T, Honda H, Seto M., "Fuzzy neural network applied to gene expression profiling for predicting the prognosis of diffuse large B-cell lymphoma", Jpn J Cancer Res. 2002 Nov;93(11):1207-12.

[3] Ressom H, Reynolds R, Varghese RS. "Increasing the efficiency of fuzzy logic-based gene expression data analysis", Physiol Genomics. 2003 Apr 16;13(2):107-17.

[4] Perez-Iratxeta C., Bork P., Andrade M.A., "Association of genes to genetically inherited diseases using data mining", Nature Genetics, vol. 31, pp. 316-319, July 2002.

[5] Ho SY, Hsieh CH, Chen HM, Huang HL. Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis. *Biosystems*. 2006 Feb 18.

[6] George J. Klir, G.J., and Bo Yuan, B., Fuzzy Sets and Fuzzy Logic: Theory and Applications, Prentice Hall Ptr., 1995.

[7] Andreasen T., Bulskov H., Knappe R., "On ontology-based querying", pp. 53-59 in Heiner Stuckenschmidt (Eds.): 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9, 2003.

[8] Tho et al., IEEE Trans. KDE, vol.18, no 6, 2006.

[9] Lee et al., IEEE Trans. SMC-B, vol.35, no 5, 2005.

[10] Lord P.W., Stevens R.D., Brass A., Goble C.A., "Semantic similarity measure as a tool for exploring the gene ontology", Pacific Symposium on Biocomputing, pp. 601-612, 2003.

[11] Speer N., Spieth C., and Zell A.. (2004) "A Memetic Clustering Algorithm for the Functional Partition of Genes Based on the Gene Ontology", Proc. of the 2004 IEEE Symposium on Comp. Intell. in Bioinf. and Comp. Biology (CIBCB 2004), San Diego, California, USA.

[12] Wang,H., Azuaje,F., Bodenreider,O., and Dopazo,J. (2004) Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In Proc. of IEEE 2004 Symposium on Computational Intelligence in Bioinformatics and Computational Biology, La Jolla, CA, USA, 25-31.

[13] Popescu M., Keller J.M., Mitchell J.A., "Fuzzy Measures on the Gene Ontology for Gene Product Similarity", IEEE Trans. Computational Biology and Bioinformatics, accepted for publication 2005.

[14] Wolstencroft, K., McEntire, R., Stevens, R., Tabanero, L. and Brass, A. (2005). Constructing ontology-driven protein family databases, Bioinformatics, Vol. 21, No. 8, pp. 1685–1692.

[15] Stevens, R., Goble, C.A. and Bechhofer, S. (2000). Ontology-based Knowledge Representation for Bioinformatics. Briefings in Bioinformatics, 1(4), pp. 398-416.

[16] Tomfohr, J., Lu, J. and Kepler, T.B., Pathway level analysis of gene expression using singular value decomposition, *BMC Bioinformatics*, 2005, 6:225.

[17] Doniger, S.W., Salomonis, N., Dahlquist, K.D., Vranizan, K., Lawlor, S.C., Conklin, B.R. (2003). MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biology 4(1):R7

[18] Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P. and Mesirov, J.P., GenePattern 2.0, *Nat. Genet.* 38 no. 5 (2006): pp500-501.

[19] Jiang J.J., Conrath D.W., "Semantic Similarity Based on Corpus Statistics and Lexical Ontology", Proc. of Int. Conf. Research on Comp. Linguistics X, 1997, Taiwan.

[20] Resnick P., "Semantic similarity in a taxonomy: an information-base measure and its application to problems of ambiguity in natural language", Journal of Artificial Intelligence Research (JAIR), 11, pp. 95-130, 1999.

[21] Yager, R., "On ordered weighted averaging aggregation operators in multicriteria decision making", IEEE Trans. On Systems, Man, and Cybernetics, 18(1),183-190, 1988.

[22] Khan S., Situ G., Decker K., Schmidt CJ, "GoFigure: Automated Gene Ontology annotation", Bioinformatics, vol. 19, no. 18, 2003.

[23] Perez AJ, Thode G., Trelles O, "AnaGram: protein function assignment", Bioinformatics, vol. 20, no.2, 2004.

[24] Prlic A., Domingues FS, Lackner P, Sippl MJ, "WILMA-automated annotation of protein sequences", Bioinformatics, vol. 20, no. 1, 2004.

[25] Bezdek, J.C., and Hathaway, R.J. (2002). VAT: A Tool for Visual Assessment of (Cluster) Tendency, Proc. IJCNN 2002, pp. 2225-2230.

[26] Theodoridis, S., and Koutroumbas, K. (2003). Pattern Recognition, 2nd Ed. New York, NY: Elsevier Science.

[27] Bezdek JC. Pattern Recognition with Fuzzy Objective Function Algorithms. New York, 1981.

[28] Popescu M., Keller J.M., Mitchell J.A., Bezdek J.C., "Functional Summarization of Gene Product Clusters Using Gene Ontology Similarity Measures", in Proc. 2004 ISSNIP, eds. M. Palaniswami,B. Krishnmachari, A. Sowmya and S. Challa, IEEE Press, pp. 553-559, Piscataway, NJ, 2004.