

Utilizing Genes Functional Classification in Microarray Data Analysis: a Hybrid Clustering Approach

George Potamias^{1,2}

¹ Institute of Computer Science, Foundation for Research and Technology – Hellas (FORTH)
71110 Heraklion, Crete, Greece

² Department of Computer Science, University of Crete
71409 Heraklion, Crete, Greece
Email: potamias@ics.forth.gr

Abstract. We present an integrated clinico-genomics environment. The proposed reference architecture provides for the seamless integration of clinical and genomic information, and aims towards the future genetic-medicine environment. Intelligent processing operations (i.e., data mining) are in the heart of this environment. In this context, we also present a novel graph-theoretic hybrid clustering approach that utilizes information about the functional classification of genes in order to achieve more knowledgeable, and by though, more naturally interpretable clustering arrangement of the genes. The clustering approach was tested on an indicative real-world datasets with satisfactory and interpretable results.

1 Introduction

The completion of DNA sequences for various organisms re-orient the related R&D agenda from static structural genomics activities to dynamic functional genomics ones. In this context *microarray* technology offers a promising alternative towards the understanding of the underlying genome mechanisms [22]. With the recent advances in microarray technology, the potential for molecular diagnostic and prognostic tools seems to come in reality. In such an integrated environment, the need to extend the standard clinical decision-making references to reliable genomic establishments raises as a major demand [3, 14].

In this paper we present the architecture and the needed operational infrastructure of an integrated environment where, clinical and genomic knowledge are appropriately fused. The aim is the accomplishment of an integrated *clinico-genomics* operational framework targeting the future *individualized genetic medicine* environment. In this respect, functional genomics and disease compacting research (e.g., inquires for disease gene-markers) are coupled and guided by related medical knowledge. An endeavor founded on the synergy between *Medical Informatics* and *Bioinformatics* [11], and oriented around the promising *microarrays* technology.

In this context the need to devise, develop and test flexible and reliable techniques for the analysis of microarray or, gene-expression data is raised [2]. *Gene-expression* data analysis is heavily depended on Gene Expression Data Mining (GEDM) technology, and in the very-last years a lot of related research efforts are in progress.

GEDM is used to identify intrinsic patterns and relationships in gene expression data (a brief outline of gene-expression data and how they are set is given in section 3.1).

The identification of patterns in complex gene expression datasets provides two benefits: (i) generation of insight into gene regulation, and (ii) characterization of multiple gene expression profiles in complex biological processes, e.g. pathological states [23].

GEDM activities are based on two approaches: (a) *hypothesis testing*- to investigate the induction or perturbation of a biological process that leads to predicted results, and (b) *knowledge discovery*- to detect ‘internal structure’ in biological data.

In this paper we present an integrated methodology that combines both. It is based on a *hybrid graph-theoretic clustering* approach able to compute and utilize different distances (or, similarities) between the objects to be clustered. In this respect the whole exploratory data analysis process becomes more *knowledgeable* in the sense that pre-established domain-knowledge is used to guide clustering.

Next section presents a proposal for an integrated clinico-genomics environment, with a focus on the ways that gene-expression profiling is adapted around the clinical decision making context. Section three, introduces the hybrid clustering approach and its utilization in the context of microarray data analysis aspects. In section four, experimental results a real-world gene-expression case study are presented. Last section concludes and presents topics for the future R&D agenda of the field.

2 Towards and Integrated Clinico-Genomics Environment

Bioinformatics enables us to understand the fundamental knowledge about biological processes. At the same time, it becomes evident that in order to fully grasp the mechanisms of a disease we do not only need an understanding of the genetic base of the disease- dealing with large amounts of data and related functional genomics approaches (such as gene-expression profiling), but we also need to *integrate the knowledge normally processed in the clinical setting*. In other words the respective R&D agenda should be forwarded towards the delivery of an Integrated Clinico-Genomics Environment (ICGE) with the combined genetic and individualized medicine being the target [14].

The Building Blocks

We envisage seven basic building blocks for the realization of the proposed ICGE (see Figure 1).

1. A set of clinical information systems to keep patients’ *clinical* information: clinical, laboratory and patholoanatomical information systems. These systems should be coupled with an information system to keep patients’ *genomic* information, i.e., the genomic information system.
2. A middleware layer for information and data integration and intelligent process.
3. Seamless and efficient *extraction* of data from the various data and information sources (clinical and genomic).

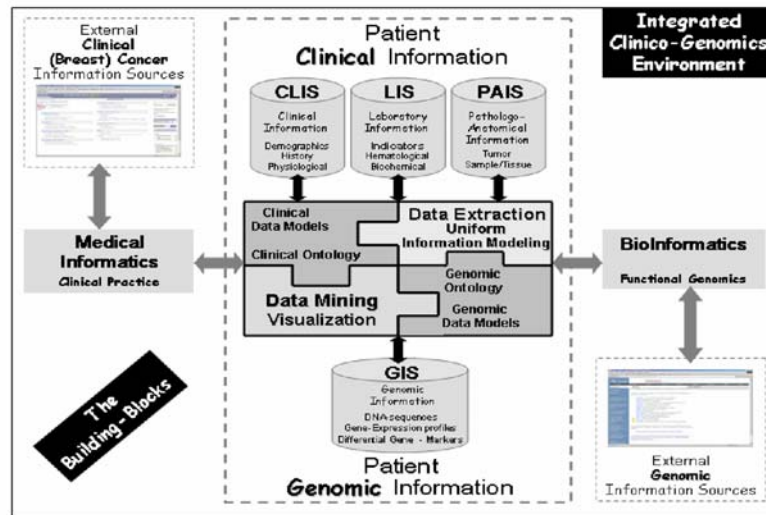


Fig. 1. The integrated Clinico-Genomics environment: Reference Architecture & Operational Components

4. *Uniform information modeling*- enabled by the utilization of standard clinical and genomic data models, as well as respective ontologies- the COAS (Clinical Object Access Service, by CORBAmed [4], HL7 [10], UMLS [25], and LOINC [12]) standard clinical interfaces and ontologies; as well as the GO (Gene Ontology, [8]), GEML, MGED and MIAME genomic/gene-expression modeling standards [7, 15, 16], offer the appropriate schemes and the needed infrastructure.
5. *Uniform representation schemes*- enabled by the utilization and the appropriate customization of RDF/XML technology.
6. *Intelligent data processing and visualization* component - enabled by a suite of *data-mining* components and tools.

The demanding clinical and genomic data integration environment post the need to elaborate on the concept of integrated electronic health care record (IEHCR) architectures [9, 24], and utilize the respective technological advances extending the standard clinical data models to genomic ones.

The Services

The following basic services should be offered by ICGE.

- * **GEDMS: GENE EXPRESSION DATA MINING SUITE:** A repository of data-mining systems and tools appropriately customized for the task of gene-expression data analysis and respective knowledge discovery operations, i.e., data-preprocessing and normalization, clustering and classification, statistical methods for gene-selection and visualization tools. The operationalization and adaptation of data-mining operations in distributed and heterogeneous information sources is a special demand [20]. The aim is to offer an environment where both molecular biologists and clinicians could easily adapt their data-analysis needs towards the identification of reliable gene-markers for diagnostic and prognostic needs.

- * ICGIS: INTEGRATED CLINICO-GENOMICS INFORMATION SERVICES: To offer the seamless access and retrieval of patients' clinical and genomic profiles. Clinical information refers to the standard patients' electronic health care record, i.e., from demographic, physiological, and historical/hereditary data to laboratory and tissue-based pathologo-anatomical findings. Genomic information refers to patients' genomic findings, i.e., dna-sequences, gene-expression profiles, gene-markers etc. The anonymity, security, authentication and authorization issues should be also addressed [19].
- * EXTERNAL INFORMATION SOURCES. A key requirement is the integration of ICGIS with *external information and knowledge sources*, as offered by clinical practice and research (e.g., sites for breast-cancer guidelines and protocols), and functional-genomics research (e.g., Human Genome Project's related Web-based information services and respective data banks).

3 Knowledgeable Gene-Expression Profiling

3.1 The Settings of Gene-Expression Profiling

By measuring transcription levels of genes in an organism under various conditions, in different tissues, we can build up *gene expression profiles*, which characterize the dynamic functioning of each gene in the genome. The gene-expression data are represented in a *matrix* with *rows* representing genes, *columns* representing samples (e.g., developmental stages, various tissues, treatments etc), and each continuous-valued cell characterizing the expression level of the particular gene in the particular sample [2].

There are two straightforward ways how gene expression matrix can be studied: (1) comparing expression profiles of genes by comparing *rows* in the expression matrix; and (2) comparing expression profiles of samples by comparing *columns* in the matrix. Additionally, both methods can be combined (provided that the data normalization allows it). When comparing rows or columns, we can look either for *similarities* or, for *differences* and accordingly form *clusters*.

Clustering: An Unsupervised Intelligent Data Analysis Approach. The goal of clustering is to group together objects (genes or samples) with *similar properties*. This can also be viewed as the reduction of the dimensionality of the system or, the discovery of 'structure in the data'. By comparing gene-expression profiles, and forming clusters, we can hypothesize that the respective genes are co-regulated and possibly functionally related.

3.2 Hybrid Clustering: Methodology Outline

A novel *hybrid clustering* approach, suitable for gene-expression data analysis, is devised. It follows three steps [21]:

- a. A *distance* is computed between all the objects (genes) to be clustered. The distance may be computed taking in consideration various *modalities*. For microarray data the distance between two genes may reflect their functional classification (i.e., their known assignment to the same or similar functional activity during the meta-

- bolic process) or, the occurrence of transcriptional-factors (i.e., pre-specified and established motifs in the corresponding DNA-sequences of the genes).
- b. A fully connected *weighted graph* is devised with the objects as nodes and the computed distances as the edges/links weights. The *minimum spanning tree* (MST) of the graph is found. The computed MST reserves the minimum distance between objects. So, objects that exhibit low distances (e.g., similar functional activities for genes or, similar expression patterns) are arranged in neighboring areas of the tree with the potential of ‘isolating’ parts of the tree and forming groups.
 - c. The MST is cut to sub-trees following an *iterative partition* process resulted into a *Graph-Theoretic Clustering* (GTC) algorithm. The final outcome is a *hierarchical* clustering-tree organization of the objects.

Coping with Sequential Data

Most of the gene-expression data are sequential, i.e., the conditions under which the expression of the different genes is studied are ordered in time. In this case we are confronted with a *time-series* clustering situation.

Discovering sequential relationships in a time sequence is important to many application domains. In data mining applications, it is often necessary to search within a series database for time-series that matches a pre-specified query series. This primitive is needed, for example, for prediction and clustering purposes. Clustering of time-series data contributes to the problem of inducing and forming categories (classes) of events. During the last years a great-deal of work is devoted on such research aspects [1, 5].

Measuring the *similarity* between objects is a crucial issue in many data retrieval and data mining applications. The typical task is to define a function $dist(a,b)$ (or, $sim(a,b)$), between two sequences a and b , which represents how ‘distant’ (or, ‘similar’) they are to each other. A simple starting point would be to measure the distance by using a normal distance metric (e.g., Euclidean). Most of the gene-expression clustering studies follow this simple approach.

But, for time-series this way of measuring distance is not appropriate, since the sequences can have outliers, different scaling factors, and baselines. As it is noted in [1], reliable time-series matching and clustering operations should take in consideration the following functions: (i) ignore small or *not-significant* parts of the series; (ii) translate the *offset* of the series in order to align them vertically; and (iii) *scale* the *amplitude* of the series so that each of the respective segments lies within an envelope of fixed width. For complex objects, designing such functions and algorithms to compute them, is by no means trivial.

Qualitative Dynamic Discretization of Time Series

The problems about identifying significant parts in time-series; offset translation and amplitude scaling could be tackled by a ‘*discretization*’ transformation of the time-series. That is, each value of a time-series is transformed into a representative *nominal* value. The nominal value depends on the discretization of the time-series values, i.e., separation of the time-series values into respective time-point intervals.

In this paper we follow and adjust the *qualitative dynamic discretization*- QDD method presented in [13] (see Figure 2). The basic idea behind this method is the use of statistical information about the preceding values observed from the series in order to select the discrete value for a new continuous value from the series. A new continuous value will be assigned to the same discrete value as its preceding values if the continuous value belongs to the same population (to be decided with a Student's t -statistic). Otherwise a static discrete transformation method (*discr* function in the lower part of Figure 2) will assign a new discrete value to this new continuous value.



Fig. 2. Time series nominalization: (a) the Qualitative dynamic discretization algorithm; and (b) computation and assignment of discrete values.

Constant patterns. With the QDT method it is very difficult to model 'constant' time-series, i.e., series with values fluctuating in 'small', i.e., non-significant ranges. We refined and enhanced the QDT method by computing a threshold value for these fluctuations.

First, each series $TS(X) = \{X_1, X_2, \dots, X_n\}$ is transformed to range in the $[0, 1]$ interval, $TS_{[0-1]}(X) = \{X_{[0-1]}, X_{[0-1],2}, \dots, X_{[0-1],n}\}$. Dividing all the values of the series by

the series' maximum value does this. Then, we use the formula below to compute the threshold value.

$$Th = \max(\{\min_{1 \leq i \leq n} X_{[0-1],i}\}) - sd(\{\min_{1 \leq i \leq n} X_{[0-1],i}\})$$

where n the total number of time-series; $\{\min_{1 \leq i \leq n} X_{[0-1],i}\}$ the list of the minimum values of all [0-1]-ranged series; $sd(\{\min_{1 \leq i \leq n} X_{[0-1],i}\})$ the standard deviation of these values.

If the minimum value of the [0-1]-ranged series transform is greater than the computed threshold then the series is considered as constant, and the discrete value s (i.e., the user specified maximum number of discrete values) is assigned to all its values (see first *if* condition in part a of Figure 2). Otherwise the discretization process is triggered (part b of Figure 2).

Notice that in Figure 2 the distance between objects is computed by matching their corresponding nominal values (the *NOM_dist* in part b of Figure 2). The current GTC implementation includes a variety of distance methods that implements: the normal and square-rooted *Euclidean* metrics; the *Pearson linear-* and *rank-correlation* metrics; the *Edit* distance metric, and the *Value Difference Metric*.

The Value Difference Metric: a Knowledgeable Distance

The VDM metric could be utilized in order to combine information about the objects that originates from different *modalities*. For example, the functional classification of genes for various organisms is known in advance (at least for organisms with completed genomes, like *yeast*) [17]. The VDM distance metric (formula below) takes into account this information, i.e., the assignment of objects to classes [27].

$$VDM_a(V_a = x, V_a = y) = \sum_{c=1}^C \left| \frac{N_{a;x;c}}{N_{a;x}} - \frac{N_{a;y;c}}{N_{a;y}} \right|^2$$

where, $V_a = x$: the value of feature a is x ; $N_{a;x}$: the number of objects value x for attribute a ; $N_{a;x;c}$: the number of objects value x for attribute a and assigned to class c ; and C : the total number of classes.

Assuming that the assignment of objects to classes is known in advance then, the VDM metric is used to utilize information that comes from external (to the feature-based description of the objects) modality. Now the clustering is to be performed on a (potentially) different distance-based arrangement of the objects, and the final hierarchical clustering outcome reflects both: (a) the feature-based description of the objects, and (b) their class assignments; in a way that the one is used to confirm/reject hypotheses about the other. So, pre-established *domain-knowledge* is utilized in order to discover regularities and confirm/reject hypotheses. In that sense, the presented clustering process presents a *knowledgeable* exploratory data analysis approach.

Graph Theoretic Clustering

Having on our disposal two different sources of information, (a) a set of discretised time-series (i.e., their feature-based description), and (b) a matrix comprising the distances between the series, the question is how we utilize both of them in order to form a reliable clustering of the series. The problem could be generalized to different kind of objects, other than time-series, and its statement has as follows:

Given:

- (i) A fully-connected weighted graph, $G(V,E)$, with each node in V representing an object, and each weighted link in E , representing the distance between the linked objects.
- (ii) A feature-based description of the objects (in our case the feature-based description refers to the discretised representation of time-series).

Find: A clustering of the objects that utilizes both (i) and (ii).

In other words we are confronted with the problem of inventing and forming categories of objects with information coming from different *modalities*, i.e., from distances and from feature-based descriptions of the objects. Towards this end, we elaborate on an innovative *graph-theoretic clustering* – GTC approach, realized within the following two steps.

1. *Minimum Spanning Tree (MST) construction.* Given a set E of n objects, the *minimum spanning tree*- MST of the *fully-connected weighted* graph of the objects is constructed; the MST contains exactly $(n-1)$ edges. A basic characteristic of the MST is that it keeps the shortest distances between the objects. This guarantees that objects lying in ‘close areas’ in the tree exhibit low distances. So, finding the ‘right’ cuts of the tree could result in a reliable grouping of the objects. This is the method followed by Zahn, [29]. But, Zahn’s MST-based clustering approach does not utilize the feature-based description of the objects, a *crucial* source of information for deciding where to cut (especially for ‘borderline’ cases). The method follows a ‘one-shot’ partition of the formed MST - appropriate ‘weak’ (or, ‘inconsistent’ in Zahn’s terminology) links are identified and cut; the nodes in the separated parts of the MST compose the formed clusters. Because of its ‘one-shot’ clustering approach the method could not identify special underlying relations in the data, as for example the potential of a hierarchical organization. With this in mind, we devised an *iterative* MST partition process concluding into a *hierarchical clustering* structure. Similar approaches for the clustering of gene-expression data have also been introduced [28].

2. *MST Partition*

- (i) *Binary Splitting.* For a current sub-cluster of objects (i.e., a *node* in the hierarchical tree), each of the edges in the corresponding MST is cut. A binary split of the objects is formed (i.e., split objects in two groups; note that if we have n objects then $n-1$ such splits are formed). The resulted two sub-clusters plus the sub-clusters formed so far (excluding the current node), compose a potential partition.
- (ii) *Best split.* The *Category Utility*- CU (see below), for all such $n-1$ partitions, is computed. The split that exhibits the *highest* CU is selected as the *best* partition of the objects.
- (iii) *Iteration & Stopping criterion.* Steps *i*, and *ii* are iteratively performed on all nodes of the tree following a *breadth-first* tree growing approach. The category utility of the current partition (as formed when a node is split), $CU_{current}$, is tested against the so far partition, CU_{so_far} . If $CU_{current} > CU_{so_far}$ then, the node is split, otherwise we *stop* partitioning on this part of the tree.

The final outcome is a *hierarchical clustering* of the input objects. GTC exhibits a time-complexity of $\sim O(n^2 \times F)$, with n the number of objects and F the number of

features [21]. This figure is for the worst case; in all experiments conducted, and for datasets with ~1000-5000 objects and ~10-20 features the real execution time ranges from 2 to 30 minutes (on 1.7MHz, 2G RAM PC).

Category Utility. For the computation and estimation of the utility that each set of clusters exhibits, we rely on the established and well-known *Category Utility* formula [6]. The CU metric resembles *information-theoretic* one, and it is based on the *distribution* of the objects' feature-values in a set of sub-clusters.

$$CU(G_1, G_2, \dots, G_g) = \frac{\sum_{k=1}^g p(G_k) \left[\left(\sum_i \sum_j p(A_i = V_{ij}/G_k)^2 \right) - \left(\sum_i \sum_j p(A_i = V_{ij})^2 \right) \right]}{g}$$

where,

$$p(G_k) = \frac{\# \text{objects in } G_k}{\# \text{total objects}},$$

$$p(A_i = V_{ij} / G_k) = \frac{\# \text{objects in } G_k \text{ with value } V_{ij} \text{ for feature } A_i}{\# \text{objects in } G_k}, \text{ and}$$

$$p(A_i = V_{ij}) = \frac{\# \text{objects with value } V_{ij} \text{ for feature } A_i}{\# \text{total objects}}.$$

The hybrid nature of the clustering approach. The formation of the MST is based on the distances between objects (step 1 of the algorithm). These distances may be computed on the basis of different modalities. Here is where the hybrid nature of the clustering process is introduced. Take as example a collection of genes with known functional classification, and given expression profiles. After the QDD is performed, the corresponding feature-based description of the expression-profiles is formed. We may use the VDM distance, with classes the functions assigned to the genes, to form the minimum-spanning tree. Then, for the partition of the objects and the formation of the hierarchical clustering tree the CU is computed with reference just to the distribution of feature-values (ignoring the functional classification of genes). With this scenario, we may conclude to groups of genes that have similar expression patterns and similar functions as well. Moreover, we may test various hypotheses. For example, in the case of a high CU for the final clustering then, we may hypothesize a strong-relation between the expression of the genes and their functions.

4 Experiments

We applied the presented GTC algorithm on an indicative gene-expression dataset. The specific study was selected because the functions of the genes are provided, and the dataset is relatively small so that the utility of the CTG algorithm is shown in a more coherent way making the interpretation of results more transparent.

Central Nervous System Development (CNSD). The CNSD gene-expression dataset presents the mRNA expression levels of 112 genes during rat central nervous system development, focusing on the cervical spinal cord. The development course is extended in nine stamped time intervals; E11, E13, E15, E18, E21 (embryonic days); P0, P7, and P14 (postnatal days); and A (or, P90- adult) (for the original the case-

study, dataset, and publication see [26]). Moreover, the assignment of the 112 genes to four main functional classes; divided further to fourteen sub-classes (i.e., Main_Class: Neuro-Glial-Markers with Sub_Classes: Markers, Neuro-transmitter-Metabolizing-Enzymes).

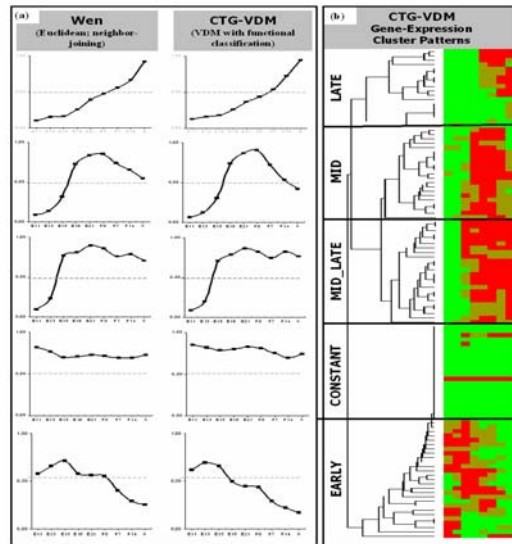


Fig. 3. Plots of the clusters' mean expression level, for Wen and GTC-VDM clustering. Gene expression patterns for the GTC-VDM clustering (for three discretization intervals).

Results and Discussion

The presented clustering approach, using the *NOM_dist* metric, was applied on the data. The results are almost identical to the ones published by Wen et al., [26]. That is, five clusters are induced: *w1/c22- early*, *w2/c2112 – mid_late*, *w3/c2111 – mid*, *w4/c212 – late*, and *w5/c22 = constant* gene-expression patterns. The same applies when we use the VDM metric (GTC-VDM clustering approach; see Figures 3, and 4). This result shows that the presented clustering approach is well formed, reliable and stable, producing similar results with the standard dendrogram-based clustering approaches (as followed by Wen's clustering approach where, a joining-neighboring clustering technique is followed). Referring to the functional classification of genes (also reported in the Wen publication), and accordingly utilizing the VDM metric, we were able to induce five clusters of genes that exhibit, not only similar expression-profiles but similar functions as well. Figure 4 shows the formed minimum spanning tree and its partition to five clusters. Nodes in different shaded colors and shapes indicate the different functions of the clustered genes.

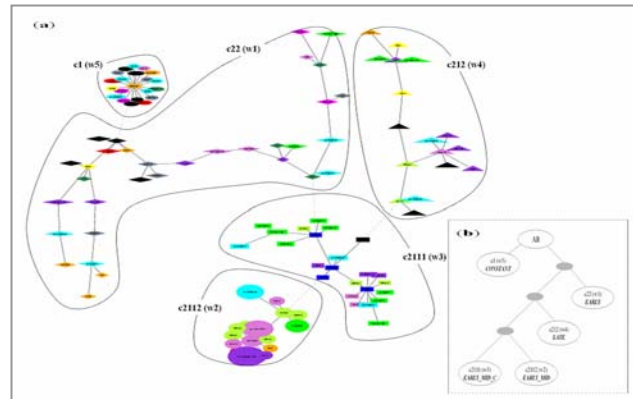


Fig. 4. (a) The MST for the CNSD domain; the five resulted sub-clusters are indicated; (b) the corresponding hierarchical clustering tree (the different colors correspond to gene functions).

To further validate the GTC-VDM clustering results, we computed an *impurity-index* (I_f) for each of the genes' function-class, relative to the induced clusters. The index is based on the distribution of the different classes into the discovered clusters, and it is computed by the following information-theoretic formula.

$$I_f = -\sum_{k=1}^C p_{f,k} \log(p_{f,k})$$

where, k the cluster index, C the total number of induced clusters, and $p_{f,k}$ the number of function-class f genes that belongs to cluster k , divided by the total number of genes assigned to function-class f .

The impurity index formula is based on a ramification of the well known *diversity index* formula [18]. It measures the degree to which the resulted clustering 'explains' the underlying function classification of the genes. If the genes of a specific function-class are equally distributed among the clusters then, the function is highly divertive (i.e., high impurity-index). In the inverse case where, the functional-classes are unequally distributed among the clusters we may conclude that the induced gene-expression cluster profiles follow the corresponding functional-classification of the genes. In other words, impurity-index helps to give answers to questions like: "do genes assigned to the 'neuro_glial_markers' function-class follow a mid_late gene expression profile?"

In Table 1 below, the impurity-indices that show a difference between the Wen and GTC-VDM clustering approaches are listed. The average index for GTC-VDM is superior (1.64 compared with 1.86 for Wen clustering). So, the GTC-VDM clustering approach induces more 'compact', with respect to the genes' functions, clusters.

One of the fundamental problems in gene-expression clustering is the right choice of the clustering parameters that influence the assignment of genes to clusters, and to the corresponding indicative expression patterns.

GENE FUNCTION	Wen	GTC-VDM
diverse_CELL_CYCLE	2.06	1.68
neurotransmitter_receptors_Glutamate	1.68	1.34
peptide_signaling_NEUROTROPHINES	2.27	2.00
diverse_INTRACELLULAR_SIGNALING	1.77	1.68
diverse_NOVEL_EST	1.68	1.48
diverse_TRANSCRIPTION_FACTOR	1.77	1.68
neuro_glial_markers_MARKERS	2.16	1.93
neurotransmitter_receptors_ACETYLCHOLINE	1.69	1.45
neurotransmitter_receptors_GABA_A	1.63	1.55
AVERAGE	1.86	1.64

Table 1. Impurity indices (2nd column) for the functions that have different distributions (columns 3-7) among the clusters of the Wen and GTC-VDM clustering approaches ('w<n>': the 'w'aves in Wen's terminology; 'c<nnn>': the GTC-VDM clusters' index number).

Especially, in hierarchical clustering approaches it is difficult to identify the 'borderline' patterns, i.e., genes with expression profiles that lie between two or more clusters. This is the situation with the w2/c2112 and w3/c2111 clusters (refer to Figure 3, above). In Wen clustering there are some genes that are assigned to cluster w2, even if their expression patterns fits more-or-less to the w3/c2111 pattern. The GTC-VDM clustering approach remedies this, and groups the genes within cluster w3/c2111.

Special cases of 'borderline' cases are the 'unclassified' ones (i.e., it is not clear to which cluster the expression-patterns of genes fit better). For example, a set of genes assigned to the 'neuro_glial_markers' function, remained unclassified with the Wen approach (the 'other' pattern in his terminology). With the GTC-VDM clustering approach most of these genes are assigned to cluster w3/c2111 in which, most of the genes are assigned to the 'neuro_glial_markers' function. So, with the utilization of *background-knowledge* (i.e., knowledge about the function of genes) it is possible to solve the 'borderline' problem, and make the interpretation of the final clustering result more natural.

5 Conclusion

We have presented an integrated clinico-genomics environment. The proposed reference architecture provides for the seamless integration of clinical and genomic information, and aims towards the future genetic-medicine environment. Intelligent processing operations (i.e., data mining) are in the heart of this environment. In this context, we presented a novel graph-theoretic hybrid clustering approach that utilizes information about the functional classification of genes in order to achieve a more knowledgeable, and by though, more naturally interpretable clustering arrangements of the genes.

The presented clustering approach is based on the discrete transformation of the gene-expression temporal profiles (a method appropriate for sequential / time-series

data), and the VDM (value difference metric) formula for the computation of distances between genes.

The approach was tested on an indicative real-world domain that refers to the large-sale gene-expression profiling of the central nervous system. The results are comparable to the original published case study Moreover, with the graph-theoretic clustering approach, and utilizing the VDM distance metric, we were able to tackle the ‘borderline’ cluster assignment problem and achieve more naturally interpretable results.

Our future research and development plans are moving towards two directions: (a) extensive and large scale experimentation with various gene-expression profiling domains, as well with other domains (i.e., time-series of economic interest), in order to test the effectiveness and the scaling-up of the approach; and (b) incorporation of the approach in the presented integrated clinico-genomics environment (an endeavor being under-way within our group in the Institute of Computer Science, FORTH, Heraklion, Crete, Greece).

References

1. Agrawal R., Lin K., Sawhney H.S., Shim K., Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. *Proceedings 21st VLDB Conference*, Zurich, Switzerland, (1995) 490-501.
2. Bassett D.E., Eisen M.B., Boguski M.S. Gene Expression Informatics: it's All in your Mind. *Nat. Genet.* 21, Supplement 1 (1999) 51-55.
3. Bioinfomed Study. Prospective Analysis of the Relationships and Synergy between Medical Informatics and Bioinformatics. <http://bioinfomed.isciii.es/> [accessed April 2003].
4. COAS, “*Clinical Observations Access Service*” (COAS), Final Submission, *OMG Document: corbamed/99-03-25*, (1999).
5. Faloutsos C., Ranganathan M., Manolopoulos Y., Fast Sequence Matching in Time-Series Databases. *Proceedings ACM SIGMOD Conference*, (1994).
6. Fisher D. Knowledge Acquisition via Incremental Conceptual Clustering. *Machine Learning*, 2: (1987)
7. GEML: Gene Expression Markup Language, Web site. <http://www.rosettahbio.com/products/conductor/geml/default.htm> [accessed April 2003].
8. GO: Gene Ontology consortium Web site. <http://www.geneontology.org> [accessed April 2003].
9. Grimson J. Delivering the Electronic Healthcare Record of the 21st century. *International Journal of Medical Informatics* 64:111–127 (2001)
10. HL7. Health Reference Information Model. <http://www.hl7.org/library/data-model/RIM/C30118/rim.htm> [accessed April 2003].
11. Kohane, I.S. Bioinformatics and Clinical Informatics: The Imperative to Collaborate. *JAMIA* 7:512–516 (2000)
12. LOINC. Logical Observation Identifiers Names and Codes. <http://www.loinc.org/> [accessed April 2003].
13. Lopez L.M., Ruiz I.F., Bueno R.M., Ruiz F.T. Dynamic Discretization of Continuous Values from Time Series. *Proceedings 11th European Conference on Machine Learning (ECML)*, Catalonia, Spain (2000) 280-291

14. Maojo V., Iakovidis I, Martín-Sánchez F, Crespo J, Kulikoswki C. Medical Informatics and Bioinformatics: European Efforts to facilitate Synergy. *Journal of Biomedical Informatics* 34(6):423-427 (2001)
15. MGED Society. <http://www.mged.org> [accessed April 2003].
16. MIAME Web site. <http://www.mged.org/Workgroups/MIAME/miame.html> [accessed April 2003].
17. Pavlidis P., Weston J., Cai J., Grundy W.N. Learning Gene Functional Classification from Multiple Data Types. *Computational Biology* 9(2):401-411 (2002)
18. Pielou E.C. The Measurement of Diversity in Different Types of Biological Collections. *Journ. Theoret. Biology*, 13:131-144 (1996)
19. Potamias G., Tsiknakis M., Katehakis D.G., Karabela K., Moustakis V., Orphanoudakis S.C. Role-Based Access to Patients Clinical Z17. Data: The InterCare Approach in the Region of Crete. *Proceedings MIE Conference*, Hannover, Germany, (2000) 1074-1079.
20. Potamias G., Moustakis V. Knowledge Discovery from Distributed Clinical Data Sources: The Era for Internet-Based Epidemiology. *Proceedings 23rd Annual International Conference IEEE Engineering in Medicine and Biology Society (EMBC)*, Instabul, Turkey (2001)
21. Potamias G. Distance and Feature-Based Clustering of Time Series: an Application on Neurophysiology. In P. Vlahavas, C.D. Spyropoulos (eds.), *Methods and Applications of Artificial Intelligence (SETN 2002)*, Springer LNAI Vol.2308 (2002) 237-248.
22. Schena M., Heller R.A., Theriault T.P., Konrad K., Lachenmeier E., Davis R.W. Microarrays: Biotechnology's Discovery Platform for Functional Genomics. *Trends Biotechnol.* 16(7):301-306 (1998)
23. Brazma A., Robinson A., Vilo J. Gene Expression Data Mining and Analysis. In: B. Jordan (ed.): *DNA Microarrays: Gene Expression Applications*, (2001) Book chapter: B 106-129.
24. Tsiknakis M., Katehakis D.G., Orphanoudakis S.C. An Open, Component-based Information Infrastructure for Integrated Health Information Networks, *International Journal of Medical Informatics*, 68(1-3):3-26 (2002)
25. UMLS. Unified Medical Language System. <http://www.nlm.nih.gov/research/umls/33> [accessed April 2003].
26. Wen X., Fuhrman S., Michaels G.S., carr, D.B., Smith S., Barker J.L., Somogyi R. Large-scale Temporal Gene Expression Mapping of Central Nervous System Development. *Proceedings Natl. Acad. Sci. (PNAS)*, (1999) 334-339.
27. D. Wilson, T. Martinez. Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research*, 6:1-34 (1997)
28. Xu Y., Olman V., Xu D. Minimum Spanning Trees for Gene Expression Data Clustering. *Genome Informatics* 12:24-33 (2001)
29. Zahn C.T. Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters. *IEEE Transactions on Computers*, 20:68-86 (1971)