Годишник на секция "Информатика"
Съюз на учените в България
Том 6, 2013, 120–133

Annual of "Informatics" Section
Union of Scientists in Bulgaria
Volume 6, 2013, 120–133

DATA INTEGRATION: AN APPROACH TO IMPROVE THE PREPROCESSING AND ANALYSIS OF GENE EXPRESSION DATA

Elena Kostadinova

Department of Computer Systems and Technologies
Technical University of Sofia – Branch Plovdiv
25 Tsanko Dyustabanov Str., 4000 Plovdiv, Bulgaria
e-mail: elli@tu-plovdiv.bg

Abstract: The integration and evaluation of data from multiple DNA microarray datasets for a specific analysis is an important and yet challenging problem. In contrast to the majority of studies, which are focused on a particular biological problem, the present paper examines how the combination of several related microarray datasets affects different areas of preprocessing and analysis of gene expression data, such as missing value imputation, gene clustering and biomarkers detection. For this purpose, three recently suggested integration models are reviewed and discussed. The biological impact of these specific integration algorithms on the three abovementioned analysis tasks is demonstrated on two types of gene expression data: time series and non-time series. The results are evaluated in terms of different validation measures.
Keywords: Microarray gene expression data, Data integration, Missing value estimation, Gene clustering, Biomarkers identification.

1. Introduction

DNA microarray technology is a powerful tool which provides the ability to monitor thousands of gene expression levels simultaneously under different conditions. All these measurements contain information about many different aspects of gene regulation and function, ranging from understanding the global cell-cycle control of microorganisms to cancer in humans. Presently, with the increasing amount of publicly available gene expression datasets, the combination of data from multiple microarray studies examining a similar biological question is gaining high importance. In general, the integration and evaluation of multiple datasets promise to yield more reliable and robust results since these results are based on a larger number of samples and the effects of individual study-specific biases are weakened. The development of methods to combine data from different microarray studies is therefore of crucial importance for their optimal use.

A series of microarray integration algorithms and meta-analysis studies have been published in the bioinformatics literature. For example, [4] and [15] perform meta-analysis of independent microarray datasets in order to identify differential expression. A method which merges multiple microarrays and computes the correlations across all datasets was presented in [7]. Zhou *et al*. addressed the microarray integration issue in [24] by proposing a technique for functional classification of genes, called second-order correlation analysis, which utilizes the pairwise correlations of gene expression across different datasets. Recently, an approach targeting the direct fusion of multi-experiment time series expression profiles was proposed in [20], and another data fusion algorithm based on multivariate regression was considered in [8]. Warnat *et al*. also addressed the direct integration issue by proposing two integration methods for deriving numerically comparable measures of gene expression from independent datasets and different microarray platforms [23]. Other integration technique that uses inter-gene information, was presented in [10].

Undoubtedly, each of the above studies has established flexible statistical model for evaluation and comparison of multiple microarray datasets. However, these papers are focused on a particular biological problem and the influence of data integration appears to be demonstrated in a rather restricted way. In contrast, the present work examines how the combination of several related gene expression datasets affects different areas of preprocessing and analysis of gene expression data. In this particular case, three types of analysis commonly seen in microarray experiments are considered: *missing value imputation*, *gene clustering* and *biomarkers detection*. For this purpose, three recently suggested integration models are reviewed and discussed. Initially, an imputation algorithm based on simultaneous analysis of multiple time series experiments is considered to demonstrate the biological impact of data integration on the missing value estimation problem. Next, a technique for deriving clustering results of multiple microarray matrices is discussed and the gene clustering performance is evaluated. Finally, the study presents an adaptive approach for integration analysis of gene dependences across different experiments and platforms. The latter is applied on a set of biologically related microarray datasets to identify differential expression of individual genes and gene pairs. The biological impact of these specific integration schemes on the three aforementioned analysis tasks is demonstrated on time series and non-time series, and the results are evaluated in terms of different validation measures.

2. Data Integration Methods

In order to demonstrate the biological impact of gene expression data integration, three recently suggested integration algorithms are presented in the following subsections.

2.1. Integrative DTW-based Imputation Algorithm

In [12], Kostadinova *et al*. addressed the problem of estimating missing values in microarray time series data using information from multiple related datasets. The rationale behind the proposed Integrative DTW-based Imputation algorithm (*IDTWimpute*) is that if a set of genes exhibit expression similarity to a gene with missing entries over multiple related datasets, then their profiles may be used in order to derive more accurate and robust estimation. The *IDTWimpute* algorithm is briefly presented below.

Assume that a particular biological phenomenon is monitored in a few high-throughput experiments under n different conditions. Thus a set of n different data matrices M_1, M_2, …, M_n will be produced, one per experiment. Each experiment is supposed to measure the gene expression levels of m genes in a number of different time points. Let M_i is a gene expression matrix that contains any missing values. In this context, the estimation process is defined as follows. Initially, all involved microarray datasets are roughly imputed in a straightforward fashion (the *first* step). Then, an appropriate set of estimation matrices is generated based on the calculated DTW distances [18] between the target matrix and the rest datasets (the *second* step). This is performed by selecting all matrices at a maximum R-DTW distance (R is preliminary defined) from the target one M_i. Suppose that a set of l estimation matrices is selected. Next, the same distance measure is applied on each estimation matrix i ($i = 1, \ldots, l$) and a corresponding quadratic DTW distance matrix dtw_i of m^2 DTW distances (one per gene pair) is built. Thus a new set of l ($l \leq n$) matrices (dtw_1, …, dtw_l) is produced. Further, a hybrid aggregation algorithm [2] is employed to transform these DTW distance matrices into a single matrix ***dtw***, consisting of one overall DTW distance per gene pair agreed between the different estimation matrices (the *third* step). Finally, a set of estimation genes all at a maximum r-overall DTW distance (r is preliminary defined) from expression profile g_j (one with missing entries) of matrix M_i is derived. The expression profiles of those genes from all estimation matrices are then used to obtain the final imputation value (the *fourth* step).

2.2. A Technique for Clustering of Multiple Microarray Datasets

In contrast to conventional clustering algorithms, where a single dataset is used to produce a clustering solution, Kostadinova *et al*. [13] developed a technique for deriving clustering results from a set of gene expression matrices. The presented method combines the information containing in multiple microarrays at the level of expression or similarity matrices and then applies a clustering algorithm on the combined matrix. Below is a brief description of the method, referred to as *ClusterIntegration*.

Assume that a particular biological phenomenon is monitored in a few high-throughput experiments under n different conditions. Each experiment i ($i = 1, 2, \ldots, n$) is supposed to measure the gene expression levels of m genes in n_i different experimental conditions or time points. Thus a set of n different data matrices M_1, M_2, …, M_n will be produced, one per experiment. In this context, the cluster integration method is defined by three distinctive steps. Initially, the set of overlapping genes is found across all the considered matrices (the *first* step). Further, some integration procedure is applied to transform the set of input matrices M_1, …, M_n into a single matrix, which values can be interpreted as consensus values supported by all the experiments (the *second* step). The latter overall matrix is then passed to the corresponding clustering algorithm (the *third* step). In [13], this idea is demonstrated by using two microarray data integration techniques (*hierarchical merge* [20] and *hybrid integration* [2]), which are specially intended to combine multiple microarrays at the level of expression or similarity matrices, and two different clustering algorithms (k-medoids and k-means). In this sense, the two data integration techniques are applied on all the input gene expression matrices and the obtained integrated similarity (or fused) matrix is passed to k-medoids (or k-means) clustering algorithm for subsequent analysis.

2.3. An Adaptive Gene Expression Data Integration Algorithm

Boeva & Kostdinova introduced in [3] a novel method for direct integration analysis of gene dependences across different laboratories, array platforms and experimental designs. The proposed algorithm, referred to as Adaptive Integration (*AIntegration*), is flexible and independent of the subsequent analysis and also, has a built-in procedure for easy adjust to newly available data. *AIntegration* is summarized in the next paragraph.

Assume that a particular biological phenomenon is monitored in a few high-throughput experiments under n different conditions. Each experiment is supposed to measure the gene expression levels of m genes in a number of different experimental conditions or time points. Thus a set of n different data matrices $M_1, M_2, \ldots, M_n$ will be produced, one per experiment. In this context, the integration algorithm consists of two phases. First, an inter-gene relation matrix G_i is constructed for each considered expression matrix M_i. Each value in the matrix G_i presents a relation (correlation) between the expression profiles of the corresponding genes. As a result, n matrices ($G_1, G_2, \ldots, G_n$) are obtained (the *first* phase). Then a recursive aggregation algorithm, discussed in [2] for time series data, is applied to transform these n matrices into a single matrix **G**, consisting of one overall interrelation value per gene pair (the *second* phase). At this stage a matrix of overall inter-gene relations obtained from previous data can be added and aggregated together with the currently constructed interrelation matrices. In this way, the previous integration results are updated with newly arriving ones studying the same phenomena. This adaptation procedure can be executed any time when new data are available. The resulting overall inter-gene relations can be considered as trade-off values agreed between the different experiments. These values express the gene correlation coefficients and therefore, may directly be analyzed to find the relationship among the genes.

3. Experimental Setup

3.1. Microarray analysis

In order to evaluate the biological impact of data integration in various microarray analysis fields, three types of analysis commonly seen in microarray experiments are considered: missing value *imputation*, *gene clustering* and *biomarkers detection*. No doubt, the three tasks remain three of the key steps in the preprocessing and analysis of gene expression data. Most of the microarray data analysis methods require complete data matrices to function and therefore, the accurate estimation of any missing entries is crucial for their optimal use. Gene clustering is a capable technique to extract meaningful information from microarray data since genes with similar expression pattern under various conditions or time points may suggest correlation and/or co-regulation in functional pathways. The differentially expressed genes are among the best characterized markers for diagnosis and prognosis of multiple diseases. In this sense, the problem of robust identification of such biomarkers from microarray data is also essential.

3.2 Datasets Description

The three considered integration algorithms (*IDTWimpute*, *ClusterIntegration* and *AIntegration*) are demonstrated and evaluated on publicly available time series and non-time series gene expression data. The time series data are obtained from a study, examining the global cell-cycle control of gene expression in fission yeast *Schizosaccharomyces pombe* [17]. The study includes nine different expression matrices (elu1, elu2, elu3, cdc25-1, cdc25-2.1, cdc25-2.2, cdc25-sep1, elu-cdc25-br, elu-cdc10-br) on which the *IDTWimpute* and *ClusterIntegration* algorithms are applied. The non-time series datasets are designed to identify a unique disease-specific gene expression that exists between end-stage Dilated cardiomyopathy (DCM) of different etiologies and non-failing (NF) human hearts [1, 11].

In order to test *AIntegration*, three different microarrays generated from two independent laboratories and experimental designs are included (Dataset A [1], Dataset B [1], Dataset C [11]). General characteristics about these non-time series datasets are summarized in Table 1.

Table 1. End-stage DCM and NF human heart based expression matrices included in the integration analysis.

Authors	Tissue origin	Microarray platform	Number of DCM samples	Number of NF samples
Barth, A. S., et al.	*Septal myocardial*	*cDNA*	*13*	*15*
Barth, A. S., et al.	*Left ventricular*	*Affymetrix U133A*	*7*	*5*
Kittleson M. M, et al.	*Left ventricular*	*Affymetrix U133A*	*21*	*6*

A special test data corpus was created in order to evaluate *IDTWimpute*. Initially, all rows containing missing values were removed from each of the nine original time series datasets. Further, the set of overlapping genes was found across the transformed original matrices and the time expression profiles of these genes were extracted. Thus, nine new matrices, referred to as *complete* (no missing values) datasets, were built. Subsequently, five new test sets were generated from each complete data matrix by deleting randomly 1%, 5%, 10%, 15%, and 20% of the data. In this way, 45 (9×5) different microarray datasets (*Experimental Data A*) were obtained to demonstrate the impact of microarray data integration on the missing value imputation performance.

The nine time series datasets are suitable for gene clustering evaluation as per the original report in [17]. Therefore, a separate test data corpus was created to assess the cluster integration method *ClusterIntegration*. Initially, the rows with more than 25% missing entries were filtered out from each of the nine expression matrix and any other missing entries were imputed by the *DTWimpute* algorithm [21]. As a result, nine complete matrices were built. Further, the set of overlapping genes was found across all datasets and their time expression profiles were extracted. Thus nine new test matrices, referred to as *Experimental*

Data B, were obtained to demonstrate the impact of microarray data integration on the gene clustering performance.

The three non-time series microarrays are suitable for differentially expressed genes detection [1, 11]. Therefore, a separate test data corpus was specially created to evaluate *AIntegration*. Initially, the expression profiles of duplicated genes in each dataset are fused by estimating their average. Further, the set of overlapping genes was found across the three original matrices and their profiles were extracted. The expression profiles are then standardized by applying z-transformation across the three datasets, since the measurements from different microarray platforms and technologies (cDNA versus Affymetrix) as well as different measurements from identical platforms cannot be put together directly in the analysis. In this way, three new non-time series matrices, referred to as *Experimental Data C*, were obtained to demonstrate the biomarker detection performance.

3.3. Experimental Methodology

3.3.1. Evaluation of Missing Value Imputation Performance

One of the goals in the present study is to demonstrate the biological impact of data integration on the missing value estimation problem. For this purpose, the performance of *IDTWimpute* is benchmarked against that of two other imputation methods - row (gene) average [19] and *DTWimpute* [21], which do not use additional information in the estimation process. In [5], de Brevern *et al.* showed that the imputation methods strongly affect the final clustering and are crucial to obtain proper clustering solutions. Therefore, the estimation performance is evaluated by studying how the three imputation methods influence on the quality of gene clustering results and which of them best succeed in restoring the correct gene correlations. The experiments are performed on the *Experimental Data A* test corpus by using two clustering algorithms (Section 3.4.1) and two different cluster validation measures for assessing the obtained clustering solutions (Section 3.4.2).

3.3.2. Evaluation of Gene Clustering Performance

The biological impact of microarray data integration on the gene clustering performance is demonstrated by using *k*-means and *k*-medoids clustering algorithms (Section 3.4.1). The latter are applied on each individual matrix of *Experimental Data B* (*i.e.* individual clustering approach) and the results are compared to those produced by applying *ClusterIntegration* algorithm on the same experimental data (*i.e.* integrated clustering approach). Recall that *ClusterIntegration* method is implemented by using two data integration techniques (*hierarchical merge* [20] and *hybrid integration* [2]), which are specially developed to combine multiple microarrays at the level of expression or similarity matrices. Since the *k*-medoids clustering algorithm requires a similarity (distance) matrix as an input dataset, the hybrid integration procedure is used to combine the quadratic similarity matrices, generated per each dataset of *Experimental Data B*. The obtained integrated similarity matrix is then passed to *k*-medoids clustering algorithm. On the other hand, *k*-means algorithm requires an original expression data matrix as input dataset and thus, the nine original matrices from the same test corpus are fused by hierarchical merge algorithm. The final fused matrix is then clustered by *k*-means algorithm. The quality of clustering

solutions in both, individual and integrated clustering approaches is evaluated by *Connectivity* and *SI* validation measures (Section 3.4.2).

3.3.3. Evaluation of Biomarkers Detection Performance

The biological impact of data integration is demonstrated by identifying genes that are differentially expressed in human DCM using multiple microarrays. For that purpose, *AIntegration* is applied on the three expression matrices of *Experimental Data C* test corpus. The following adaptive procedure is performed. First, the input data were grouped into end-stage DCM and non-failing samples. In this way, two new matrices were obtained for each considered expression matrix: one that contains all samples from the group of end-stage DCM hearts and another including all samples coming from the donor hearts. Thus two groups of expression matrices (six matrices in total) were obtained. An inter-gene relation matrix was generated for each considered expression matrix by using the Euclidean distance. Then the recursive aggregation algorithm was applied in order to transform each group of matrices into a single matrix, consisting of one overall interrelation value per gene pair. Thus two integrated matrices were produced, one per each group. Finally, these two inter-gene relation matrices were transformed into a single matrix, which contains at each position the difference between the corresponding overall interrelation values of two matrices. The latter matrix can be used to identify the most discriminative gene pairs and genes, respectively. A relevant selection algorithm, referred to *R-Radius Scoring Pairs* (RRSP), was specially developed to identify a set of pairs, which have at least R discriminative value. R is preliminary determined as a percentage of the discriminative score of the top scoring gene pair. The RRSP algorithm is applied on the individual and integrated matrices and the gene pairs (respectively, genes) with most prominent changes in expression are compared.

3.4. Clustering Algorithms

Two partitioning algorithms are commonly used for the purpose of dividing data objects into *k* disjoint clusters [14]: *k*-means clustering and *k*-medoids clustering. Both algorithms start by initializing a set of *k* cluster centers, where *k* is preliminarily determined. Then, each object of the dataset is assigned to the cluster whose center is the nearest, and the cluster centers are recomputed. This process is repeated until the objects inside every cluster become as close to the center as possible and no further object item reassignments take place. The expectation-maximization (EM) algorithm [6] is commonly used for that purpose, *i.e.* to find the optimal partitioning into *k* groups. The two partitioning methods in question differ in how the cluster center is defined. In *k*-means clustering, the cluster center is defined as the mean data vector averaged over all objects in the cluster. For *k*-medoids clustering, which is a robust version of the *k*-means, the cluster center is defined as the object which has the smallest sum of distances to the other objects in the cluster, *i.e.* this is the most centrally located point in a given cluster.

The partitioning algorithms contain the number of clusters (*k*) as a parameter and their major drawback is the lack of prior knowledge for that number to construct. Unfortunately, determining a correct, or even suitable, *k* is a difficult, if not impossible, problem in a real microarray dataset. For such cases, researchers usually try to generate clustering results for a range of different numbers of clusters and subsequently assess the quality of the obtained clustering solutions. In a previous work [13], *k*-means and *k*-medoids clustering algorithms

were studied on fission yeast *Schizosaccharomyces pombe* data [17] for different values of k and the values between 5 and 15 were found to generate good clustering solutions. Therefore a value in this range was used for the performed experiments.

3.5. Cluster Validation Measures

Silhouette Index [16] is a cluster validity index that reflects the compactness and separation properties of any clustering solution (partition) C_1, C_2, …, C_k. Suppose a_i represents the average distance of gene i to the other genes of the cluster to which the gene is assigned, and b_i represents the minimum of the average distances of gene i to genes of the other clusters. Then the *Silhouette Index* (*SI*) of matrix M, which contains the expression profiles of m genes, is defined as $s(k, M) = 1/m \cdot \sum_{i=1}^{m} (b_i - a_i) / \max\{a_i, b_i\}$. The values of *SI* vary from -1 to 1 and higher value indicates better clustering results.

Connectivity captures the degree to which genes are connected within a cluster by keeping track of whether the neighboring genes are put into the same cluster [9]. Therefore, it is used to assess the connectedness property of clusters. Define $m_{i(j)}$ as the jth nearest neighbour of gene i, and let $x_{im_{i(j)}}$ be zero if i and j are in the same cluster and $1/j$ otherwise. Then for a particular clustering solution C_1, C_2,…, C_k of matrix M, which contains the expression values of m genes in n different experimental conditions or time points, the connectivity is define as $Conn(k, M) = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{im_{i(j)}}$. The connectivity has a value between zero and infinity and should be minimized.

4. Results and Discussion

4.1. Evaluation of Missing Value Imputation Performance

To demonstrate the biological impact of data integration on the missing value imputation performance, the *IDTWimpute* algorithm which uses information from multiple related datasets is benchmarked against two other algorithms that are based on a single microarray matrix. Fig. 1 depicts the *SI* values generated by applying the k-medoids algorithm on the imputed *Experimental Data A* test matrices with 1%, 5%, 10%, 15% and 20% missing values estimated by using row (gene) average (*RowAverage*), *DTWimpute* and *IDTWimpute*, respectively. The clustering method was run for $k = 7$ number of clusters. As it can be seen in Fig. 1, the *IDTWimpute* estimation algorithm outperforms *RowAverage* for almost all the datasets and produces statistically similar *SI* results to those obtained by *DTWimpute* for low missing rates (1%, 5% and 10%). However, the performance of *IDTWimpute* can be clearly seen for a higher missing value percentage (15% and 20%) since the latter algorithm is superior in most of the experiments. This is may be due to the fact that the integrative imputation method under question manages more adequately to preserve the gene correlation structure of the estimated data than the imputation methods based solely on a single expression matrix. The same behavior is demonstrated by using k-means clustering algorithm. The impact of the three missing value imputation algorithms is further investigated by assessing the connectedness property of produced clustering solutions in terms of *Connectivity* scores. The conducted experiments support the results generated by the *SI*

validation index (results not shown). Therefore, the imputation approach which uses integrated information may be considered as a more robust solution with respect to the quality of gene clustering than the conventional missing value estimation algorithms. This will be useful for estimating data sets with a limited number of samples, where the information is not sufficient to select neighbor genes accurately.

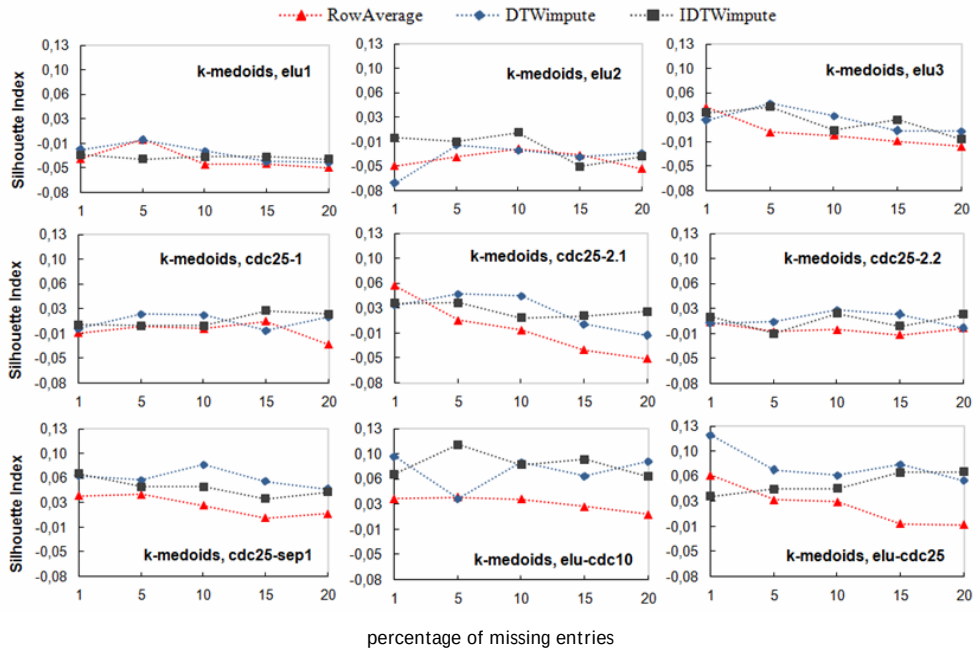


Figure 1. *SI* values generated by applying the *k*-medoids algorithm on the imputed test matrices with 1%, 5%, 10%, 15% and 20% missing values estimated by the *RowAverage*, *DTWimpute* and *IDTWimpute*, respectively.

4.2 Evaluation of Gene Clustering Performance

The biological impact of data integration on the gene clustering performance is demonstrated by applying *k*-means and *k*-medoids clustering algorithms on the individual and integrated gene expression datasets. Next Fig. 2 presents the *SI* (2a-2b) and *Connectivity* (2c-2d) values generated on each individual *Experimental Data B* test dataset and on the combined by *ClusterIntegration* expression and distance matrices from the same test corpus. In addition, the *SI* and *Connectivity* scores that are produced on the matrix obtained by averaging the corresponding values of the individual distance matrices are depicted in Fig. 2b and Fig. 2d. The clustering methods were studied for two different cluster numbers $k = 7$, 10. It can be seen in Fig. 2a and Fig. 2c that the *SI* (respectively, *Connectivity*) scores produced on the fused expression matrix are in most cases (especially, for $k = 7$) better than

those obtained on the individual ones. Moreover, they are comparable to those given on the best performing individual dataset (elu-cdc10). In contrast, the results obtained on the individual datasets outperform the results generated on the integrated distance matrix (Fig. 2b and Fig. 2d). The latter may be due to the fact that the values in the integrated distance matrix are calculated by applying the hybrid aggregation algorithm, which produces trade-off values agreed between the individual distance matrices. These compromised distance values certainly affect the performance of *k*-medoids algorithm with respect to cluster separation and connectedness. However, the integrated distance matrix performs better than the corresponding averaged matrix for both values of *k* in terms of *SI* and *Connectivity* validation indices.

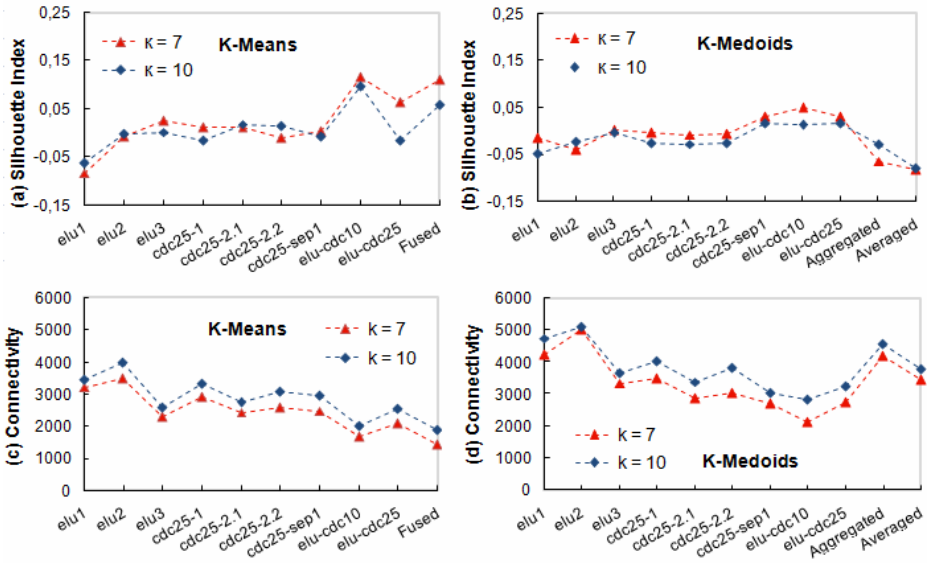


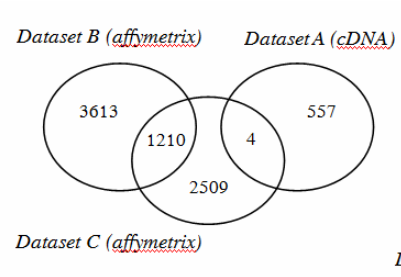
Figure 2. *SI* (a)-(b) and *Connectivity* (c)-(d) results obtained on individual datasets versus the corresponding values generated on the fused expression (integrated distance, respectively) matrix.

4.3. Evaluation of Biomarkers Detection Performance

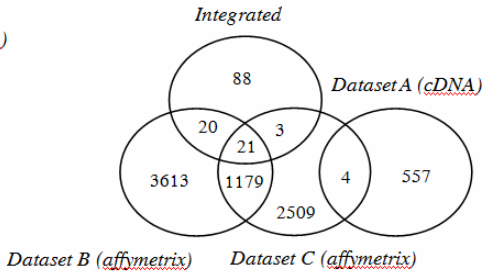
The biological impact of data integration is demonstrated by identifying marker genes for DCM detection using integration analysis of three datasets, coming from two different microarray platforms (cDNA vs. Affymetrix). First, the three gene expression matrices of *Experimental Data C* test corpus were integrated by *AIntegration* algorithm. Then, the RRSP algorithm was applied on the individual and integrated matrices and the gene pairs with most prominent changes in expression for each individual matrix and for the integrated matrix were selected. The results were obtained for R = 65% of the discriminative value of the top scoring gene pair. The Venn's diagram in Fig. 3a depicts the number of selected gene pairs on each individual matrix and the number of consistently identified across each pair of matrices. As one can see, the agreement between the datasets coming from two different

array platforms (cDNA vs. Affymetrix) is remarkably poor. Namely, only a very small number of 4 common marker gene pairs between experimental data A and C (versus none shared in A and B) were identified. In contrast to this lack of consistency, Fig. 3b shows that the proposed integration method can always identify a set of marker gene pairs, which are supported by all the experiments. Thus, though the three individual matrices have an empty intersection, 88 most discriminative gene pairs were found on the base of integration analysis of the individual experiments. As one can notice in Fig. 3b the half of these gene pairs were also identified as bio-markers either by dataset B either by dataset C, but only a quarter of the selected pairs are supported by both experiments.

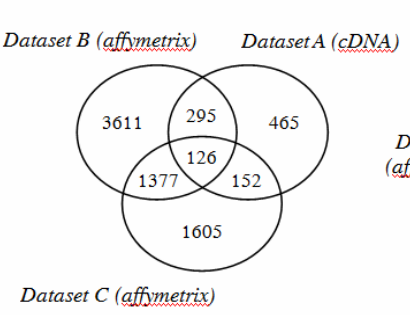
Further, the Venn's diagram in Fig. 3c shows the number of selected individual genes on each individual matrix, the number consistently identified across each pair of matrices and the number identified on all three datasets. In this context, the intersection is not empty and it contains 126 differentially expressed genes. Fig. 3d further extends the considerations with the integrated matrix.



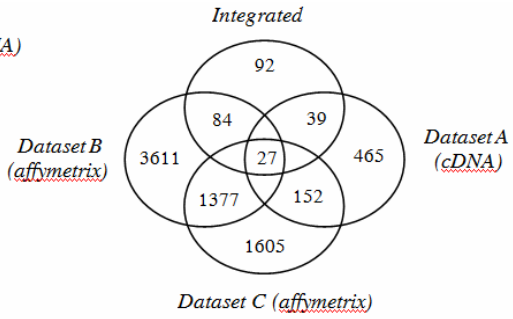
(a) Overlap of gene pairs between the individual matrices



(b) Overlap of gene pairs between the individual and integrated matrices



(c) Overlap of genes between the individual matrices



(d) Overlap of genes between the individual and integrated matrices

Figure 3. The Venn diagrams illustrate the number of identified R-radius scoring gene pairs (individual genes) in human DCM on the individual and integrated matrices.

It can again be observed that only a quarter of the genes selected on the base of the integration analysis (92 genes) are as well identified simultaneously by the individual experiments (27 genes). The potential of *AIntegration* was further investigated by comparing the list of selected individual genes for R = 80% of the top scoring gene pair's discriminative value to those identified by applying the well-established statistical procedure Significance Analysis of Microarray (SAM) [22] on the data of Barth, *et al.* [1] (Datasets A & B). SAM is intended to determine whether the changes in gene expression are significant by assimilating a set of gene-specific test statistics. It was found that 87.5% (7 of 8 identified genes) of the genes selected by the integrated matrix are in fact not presented in the intersection of the gene's lists, obtained by single-set analysis performed in [1] (Table 2). Evidently, the identification of discriminative gene expression signatures of DCM across different microarray platforms and independent studies may significantly be improved by applying the integration approach, rather than relying only on the intersection of the single-set results. In addition, once the marker gene pairs are selected, a second-order gene analysis can be performed by additionally studying the most frequently appearing individual genes, *i.e.* those that participate in more gene pairs. For instance, gene NPPB from Table 2 takes part in the top-scoring gene pair and, in addition, it is paired with other five genes. Such genes may be assigned a higher importance by the classification algorithm.

Table 2. The most discriminative genes identified by *AIntegration* versus those selected by SAM analysis.

Gene Name	Dataset A (SAM)	Dataset B (SAM)	Intersection of A and B (SAM)	Integrated Matrix
ACTA1, CKM, CRYAB, MB, MYL2, RPL23A, TNNI3	+	-	-	+
NPPB	+	+	+	+

5. Conclusion

The main contribution in this paper is to evaluate and demonstrate the biological impact of microarray data integration on three of the key steps in the preprocessing and analysis of gene expression data: missing value imputation, gene clustering and biomarkers detection. For this purpose, three different integration algorithms (*IDTWimpute*, *ClusterIntegration* and *AIntegration*) are reviewed and discussed. The imputation algorithm based on simultaneous analysis of multiple time series experiments *IDTWimpute* is shown to preserve better the gene correlation structure of the estimated data than the conventional estimation algorithms, based solely on a single matrix. The technique to derive clustering results of multiple microarray matrices *ClusterIntegration* is demonstrated to yield better clustering solutions, compared to those obtained on the individual datasets. Finally, the adaptive algorithm for

integration analysis of gene dependences across different experiments and platforms *AIntegration* is shown to be a robust and flexible procedure for DCM marker genes identification. The impact of these specific integration algorithms is demonstrated with time series and non-time series gene expression data. Although several prior studies have investigated the impact of data integration on the abovementioned types of analysis, to the author's knowledge this paper is the first to systematically evaluate the impact of data integration on all the three microarray analysis fields.

References

[1] Barth, A.S., *et al.*, Identification of a Common Gene Expression Signature in Dilated Cardiomyopathy Across Independent Microarray Studies, *Journal of the American College of Cardiology*, vol. 48, No. 8, 2006, 1610–1617.

[2] Boeva, V., E. Kostadinova, A Hybrid DTW based Method for Integration Analysis of Time Series Data, *Proc. in ICAIS'09*, Austria, 2009, 49–54.

[3] Boeva, V., E. Kostdinova, An Adaptive Approach for Integration Analysis of Multiple Gene Expression Datasets, *AIMSA'2010, LNAI 6304*, Springer-Verlag Berlin Heidelberg, 2010, 221–230.

[4] Choi, J.K., *et al*., Combining multiple microarray studies and modeling interstudy variation, *Bioinformatics*, vol. 19, 2003, pp. i84-i90.

[5] de Brevern, A.G., *et al.*, Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering, *BMC Bioinformatics*, Vol. 5, 2004, pp. 114.

[6] Dempster, A.P., N.M. Laird, D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, B*, Vol. 39(1), 1977, 1–38.

[7] Eisen, M., P. Spellman, P. Brown, D. Bostein, Cluster analysis and display of genome-wide expression patterns, *PNAS*, Vol. 95(25), 1998, 14863–14868.

[8] Gilks, W. R., B. D. M. Tom, A. Brazma, Fusing microarray experiments with multivariate regression, *Bioinformatics*, Vol. 21(2), 2005, ii137–ii143.

[9] Handl, J., *et al.*, Computational cluster validation in post-genomic data analysis, *Bioinformatics*, Vol. 21, 2005, 3201–3212.

[10] Kang, J. *et al*., Integrating heterogeneous microarray data sources using correlation signatures, *LNBI* 3615, Springer-Verlag Berlin Heidelberg, 2005, 105–120.

[11] Kittleson, M.M., K.M. Minhas, R.A. Irizarry, *et al.*, Gene expression analysis of ischemic and nonischemic cardiomyopathy: shared and distinct genes in the development of heart failure, *Physiological Genomics*, Vol. 21, 2005, 299–307.

[12] Kostadinova, E., V. Boeva, L. Boneva, and E. Tsiporkova, An Integrative DTW-based Imputation Method for Gene Expression Time Series Data, *The 6th IEEE*

International Conference on Intelligent Systems IS'12, Sofia, Bulgaria, 2012, 258–263.

[13] Kostadinova, E., V. Boeva, and N. Lavesson, Clustering of Multiple Microarray Experiments Using Information Integration, C.Bohm *et al*. (Eds.): *ITBAM'2011,* LNCS 6865 Springer-Verlag Berlin Heidelberg, 2011, 123–137.

[14] MacQueen, J.B., Some methods for classification and analysis of multivariate observations, *Proc. Fifth Berkeley Symp. Math. Stat. Prob*., Vol. 1, 1967, 281–297.

[15] Rhodes, D.R., *et al.*, Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression, Proc. Natl. Acad. Sci. USA, Vol. 101, 2004, 9309–9314.

[16] Rousseeuw, P., Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational Applied Mathematics*, Vol. 20, 1987, 53–65.

[17] Rustici, G., *et al.*, Periodic gene expression program of the fission yeast cell cycle, *Nat. Genetics*, Vol. 36, 2004, 809–817.

[18] Sakoe, H. and S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, *IEEE Trans. on Acoust, Speech, and Signal Proc*., Vol. ASSP-26, 1978, 43–49.

[19] Troyanskaya, O., *et al.*, Missing value estimation methods for DNA microarrays, *Bioinformatics*, Vol. 17, 2001, 520–525.

[20] Tsiporkova, E. and V. Boeva, Fusing Time Series Expression Data through Hybrid Aggregation and Hierarchical Merge, *Bioinformatics*, Vol. 24(16), 2008, i63–i69.

[21] Tsiporkova, E. and V. Boeva, Two-pass imputation algorithm for missing value estimation in gene expression time series, *JBCB*, Vol. 5, 2007, No. 5, 1005–1022.

[22] Tusher, V.G., R. Tibshirani, G. Chu G, Significance analysis of microarrays applied to the ionizing radiation response, *Proc Natl Acad Sci U S A*, Vol. 98, 2001, 5116–5121.

[23] Warnat, P., R. Eils, and B. Brors, Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes, *BMC Bioinformatics*, Vol. 6, 2005. No. 1, p. 265.

[24] Zhou, X.J., *et al.*, Functional annotation and network reconstruction through cross-platform integration of microarray data, *Nature Biotechnology*, Vol. 23(2), 2005, 238–43.