IRIS AperTO

UNIVERSITÀ DEGLI STUDI DI TORINO

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

Leveraging additional knowledge to support coherent bicluster discovery in gene expression data

This is the author's manuscript

Original Citation:

Availability:

This version is available http://hdl.handle.net/2318/148701 since 2026-07-04T12:57:16Z

Published version:

DOI:10.3233/IDA-140671

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

10 October 2026

UNIVERSITÀ DEGLI STUDI DI TORINO

This is an author version of the contribution published on:

A. Visconti, F. Cordero, R.G. Pensa
Leveraging additional knowledge to support coherent bicluster discovery in gene expression data
INTELLIGENT DATA ANALYSIS (2014) 18
DOI: 10.3233/IDA-140671

The definitive version is available at:
http://iospress.metapress.com/content/838k072521037162/?p=2738da488e6d48b5ae2243903ef32c24π=4

Leveraging additional knowledge to support coherent bicluster discovery in gene expression data

Alessia Visconti, Francesca Cordero, and Ruggero G. Pensa
Department of Computer Science, University of Torino, Italy
ruggero.pensa@unito.it

Abstract

The increasing availability of gene expression data has encouraged the development of purposely-built intelligent data analysis techniques. Grouping genes characterized by similar expression patterns is a widespread accepted – and often mandatory – analysis step. Despite the fact that a number of biclustering methods have been developed to discover clusters of genes exhibiting a similar expression profile under a subgroup of experimental conditions, approaches driven by similarity measures based on expression profiles alone may lead to groups that are biologically meaningless. The integration of additional information, such as functional annotations, into biclustering algorithms can instead provide an effective support for identifying meaningful gene associations.

In this paper we propose a new biclustering approach called Additional Information Driven Iterative Signature Algorithm, AID-ISA. It supports the extraction of biologically relevant biclusters by leveraging additional knowledge. We show that AID-ISA allows the discovery of coherent biclusters in baker's yeast and human gene expression data sets.

1 Introduction

The analysis of gene expression data is a crucial step for a wide variety of scientific studies involving, among the others, regulatory network inference [26], drug discovery [38], biomarkers identification [17], cell differentiation analysis [8], phylogenetics [12], and so on.

Gene expression data is usually represented with matrices whose entries are expression values of a set of genes (the rows of the matrix) in a set of biological conditions of interest (the columns of the matrix).

Clustering techniques have been commonly used in data analysis. The term *clustering* refers to a class of algorithms that partition data into groups so to maximize the intra-group similarity and minimize the inter-group similarity [21]. In the context of gene expression analysis, clustering algorithms enable the discovery of homogeneous gene (or experimental condition/sample) groups based on their expression profiles [13]. A limitation of a traditional cluster analysis is that it involves metrics taking into account expression values in *all* conditions. For instance, given a matrix of m genes and n samples, the commonly adopted

Euclidean distance between two genes is computed on two vectors involving all n expression values. However, in most cases, co-expressed genes are such only in a subset of experimental conditions. Hence, there is a strong correlation between the set of co-expressed genes and the set of experimental conditions (or biological samples) in which this co-expression is observed. For instance, a group of genes may show a similar expression profile as a metabolic response to a drug treatment concerning only a subset of biological samples. Alternatively, a group of genes may be co-expressed because they characterize a histological type of cancer samples.

To exceed this limitation, biclustering approaches have been proposed [25]. They examine gene and sample dimensions simultaneously, enabling the discovery of coherent and meaningful biclusters, i.e., potentially overlapping groups of genes showing similar activity patterns under a specific subset of experimental conditions. Biclustering has been proved to be useful for revealing potential transcriptional modules, i.e., subsets of co-expressed — and thus co-regulated — genes and of experimental conditions presenting this co-regulation [19]. Even though applying traditional cluster analysis on genes first and on samples afterwards (or vice versa) reveals groups that are similar to biclusters, clustering genes and samples separately is different than clustering them simultaneously. In the latter case, in fact, the metric to be optimized considers necessarily the association between genes and samples.

Many biclustering methods tailored for gene expression data analysis have been developed so far. For instance, Cheng and Church define a bicluster as a maximal *genes* $\times$ *samples* submatrix having a small mean squared residue score [7]. Ihmels *et al.*, instead, propose an algorithm that starts with a random bicluster and iteratively updates it in order to maximize the homogeneity of genes and samples within the bicluster [19].

Both clustering and biclustering approaches mainly use distance metrics based only on expression levels and thus not optimized to capture biologically meaningful groups. The main reason is that expression values are often highly noisy, due to measurement errors, instrumentation defaults, and to their intrinsic susceptibility to non-controlled fluctuations involving still unknown factors. To cope with this problem, several works define original distance metrics based on multiple sources of information, such as Gene Ontology, biological networks, operon annotations, intergenic distances, and transcriptional coresponses [5, 16, 42]. These works first define several metrics from additional knowledge, then adopt classical clustering methods leveraging these metrics. Unfortunately, these approaches have been proposed for clustering only. Recently, a constrained biclustering algorithm has been proposed to combine gene expression with user-defined (biological) constraints [31], but this approach does not allow for group overlapping and it is limited to pairwise (*must-link* and *cannot-link*) and temporal constraints.

In this paper we propose a new biclustering approach, called Additional Information-Driven Iterative Signature Algorithm (AID-ISA). It uses an alternate refinement process based on additional information and combined with the Iterative Signature Algorithm [19]. When additional information on genes and/or samples is available, our approach leverages such knowledge to support the discovery of meaningful groups of genes. Differently from the above mentioned attempts, our algorithm aims at identifying coherent, overlapping, and biologically-meaningful biclusters, adopting potentially any kind of exter-

nal source of knowledge on genes and/or conditions. An extensive experimental study on two *Saccharomyces cerevisiae* microarray data sets and a human one shows that AID-ISA supports the discovery of coherent biclusters that can be comfortably associated to transcriptional modules.

The rest of the paper is organized as follows: Section 2 explores the state-of-the-art in related literature; Section 3 describes how to support a biclustering algorithm with a refinement process driven by additional sources of information; Section 4 reports the results of our experimental study on three gene expression data sets; finally, Section 5 carries out some concluding remarks about this work.

2 Related work

In the context of gene expression data analysis, several authors have considered the computation of potentially overlapping local patterns (*biclusters*, see [25] for a survey).

Cheng and Church propose the so-called *biclustering approach for gene expression data* [7]. They define a bicluster as a subset of rows and subset of columns that identifies a submatrix having a low mean-squared residue. When this measure is equal to 0, the bicluster contains rows having the same value on all the bicluster columns; when it is greater than 0, one can remove rows or columns to decrease this value. The proposed method finds maximal-size biclusters such that the mean-squared residue is lower than a given threshold. The same definition of residue has been used by Dhillon *et al.* [9]. They propose two sum-squared residue measures, showing that the one defined by Cheng and Church fits better to gene expression data analysis; then, they introduce their biclustering algorithm which optimizes these residue functions. Differently from Cheng and Church's original work, this algorithm identifies a grid of non-overlapping biclusters covering the whole gene expression matrix. Ihmels *et al.* [20, 19] propose an iterative two-step process, the so called *Iterative Signature Algorithm* (ISA), which builds a bicluster starting from a normalized gene expression matrix and from a random bicluster. Another important biclustering approach is the one introduced by Tanay *et al.*, who describe a heuristic method, called SAMBA, that combines a graph-theoretic approach with a statistical data model [43]. All these approaches consider metrics computed on the expression profiles as the sole criteria to assess each bicluster.

Designing new measures to combine different source of information is not an easy task. The pioneers of a stream of works addressing this problem were Hanisch and co-workers [16] that proposed a novel approach allowing for an entirely exploratory joint analysis of gene expression data and biological networks. They define a measure derived from gene expression values and from metrics evaluated on biological networks, which is used as distance function in a hierarchical average linkage clustering algorithm. Starting from this work, Steinhauser *et al.* [42] propose a new measure combining operon annotations, intergenic distance, and transcriptional co-response data into a distance metric used in hierarchical clustering algorithms. More recently, Brameier *et al.* [5] have presented a co-clustering approach based on self-organizing maps (SOMs), where center-based clustering of standard SOMs are combined with a representative-based clustering. The authors develop a two-level cluster selection where the nearest cluster according to a distance based on Gene Ontology [2] is selected among

the best matching clusters according to gene expression distance. In this work, co-clustering means that ontology-based clustering and expression-based clustering are performed in parallel. None of these methods performs biclustering on both genes and samples at the same time.

Another more recent way to include additional knowledge in a bicluster analysis process is the so-called *constrained biclustering* (or constrained co-clustering) approach. Pensa and Boulicaut were the first to address the problem of co-clustering under user-defined constraints [30]. They extend Cho *et al.*'s approach [9] by allowing the satisfaction of must-link, cannot-link, and temporal constraints. Also, in a subsequent work, they propose a constrained co-clustering formulation that generalizes the previous approach by exploiting Bregman divergences [31]. More recent approaches of constraint-based co-clustering have been applied to textual data [40, 37]. In these works, the authors use the term *co-clustering* to identify a class of algorithms that, similarly to the work of Cho *et al.*, build a grid of non-overlapping biclusters that they call co-clusters.

Item sets and association rules have also been used for the extraction of putative transcriptional modules from gene expression data [3]. Besson *et al.* describe an algorithm for mining closed item sets, i.e., maximal biclusters, in 0/1 matrices that also embed additional knowledge related to transcription factors [4]. Unfortunately, these techniques require a non-trivial pre-processing step to discretize the gene expression values; moreover they usually end up with thousands of redundant biclusters whose post-processing demands an important effort.

In this work, we introduce an algorithm that leverages additional external knowledge by modifying consistently the way algorithm ISA [19] optimizes each bicluster. The adopted strategy is borrowed from constrained co-clustering approaches [31] in that it implicitly decides whether two genes/samples are candidate to be linked in the same cluster or not by exploiting a distance computed on additional features.

3 Biclustering with additional knowledge

In this section we introduce AID-ISA, our algorithm for mining biclusters that leverages on additional sources of knowledge. The algorithm is based on a refinement approach, called *Additional Information-Driven* (AID) process, embedded into the well known *Iterative Signature Algorithm* (ISA). AID-ISA takes both expression profiles and additional sources of information into account to discover biologically meaningful biclusters.

Before delving into theoretical details of our approach, let us introduce some notation. Let $A \in \mathbb{R}^{m \times n}$ denote a gene expression matrix. Let a_{ij} be the expression level corresponding to the ith gene under the jth condition. Let $I \subseteq \{1, \ldots, m\}$, $|I| = k$ and $J \subseteq \{1, \ldots, n\}$, $|J| = l$ be clusters of genes and samples, respectively. A bicluster $B \in \mathbb{R}^{k \times l}$ is a submatrix of A specified by the pair (I, J), in formula: $B = \{a_{ij} | i \in I, j \in J\}$. The problem addressed by a biclustering algorithm is the identification of a set of biclusters such that each bicluster $B_h = (I_h, J_h)$ satisfies some homogeneity conditions. Let us identify, for both genes and samples, a set of features describing genes/samples themselves, i.e., the additional information. By leveraging features we define two

Algorithm 1 AID: Additional Information-Driven process for cluster refinement

Input: I, m, D^G, δ_r, δ_e
Output: I

1: $\hat{\imath} \leftarrow \underset{i \in I}{\operatorname{argmin}} \sum_{\forall i' \in I, i \neq i'} D^G_{i,i'}$
2: $\bar{d}_I \leftarrow \frac{\sum_{\forall i,i' \in I, i \neq i'} D^G_{i,i'}}{2(|I|-1)}$

3: **for all** $i \in I$ **do**
4: **if** $D^G_{i,\hat{\imath}} > \delta_r \bar{d}_I$ **then**
5: $I \leftarrow I \setminus \{i\}$

6: $\bar{d}_I \leftarrow \frac{\sum_{\forall i,i' \in I, i \neq i'} D^G_{i,i'}}{2(|I|-1)}$

7: **for all** $i' \in \{1, \ldots, m\}$, $i' \notin I$ **do**
8: **if** $D^G_{i',\hat{\imath}} \leq \delta_e \bar{d}_I$ **then**
9: $I \leftarrow I \cup \{i'\}$

distance matrices, namely $D^G \in \mathbb{R}^{m \times m}$ (the gene distance matrix) and $D^C \in \mathbb{R}^{n \times n}$ (the sample distance matrix). Each matrix entry D_{pq} is set to the distance between the pth and the qth object. Potentially any definition of distance can be adopted here, also depending on the type of additional information, that can be provided in form of numeric/boolean features, strings/sequences/graphs, images and so on. Thus, our approach is very general and may be adapted to the specific analysis task by simply choosing the right distance metrics.

3.1 Additional Information-Driven refinement process

We propose a new Additional Information-Driven refinement process. It uses the information contained in the distance matrices to adjust biclusters by refining the set of genes/samples they include. AID does not produce biclusters itself, but can be used as a general building block for biclustering algorithms.

AID is summarized in Algorithm 1. In the following we describe its application on gene dimension, but it is intended that it can be applied symmetrically on the sample dimension.

In the first step AID selects a cluster representative $\hat{\imath}$ (line 1), defined as:

$$\hat{\imath} = \underset{i \in I}{\operatorname{argmin}} \sum_{\forall i' \in I, i \neq i'} D^G_{i,i'},$$

where $D^G_{i,i'}$ is the distance value between the ith and the i'th genes. $\hat{\imath}$ is the object closest to objects belonging to I. Afterwards, the algorithm evaluates the cluster average distance $\bar{d}_I$ (line 2) as:

$$\bar{d}_I = \frac{\sum_{\forall i,i' \in I, i \neq i'} D^G_{i,i'}}{2(|I|-1)}.$$

Genes in I that are distant from the representative more than δ_r times the average cluster distance $\bar{d}_I$ are removed from the cluster (line 3-5). Then, the cluster average distance is updated to reflect the change in the bicluster composition (line 6). Finally, all genes not belonging to I having a distance from $\hat{\imath}$ smaller or equal than δ_e times the new average cluster distance, are added to I (line 7-9). This step includes in the bicluster objects that are likely to be related, but that have been excluded by the main biclustering process so far.

3.2 Additional Information-Driven ISA

We modify the Iterative Signature Algorithm (ISA) [19] to embed the AID refinement process. ISA is a two-step iterative procedure. It starts from a random set of genes to which a default score of 1 is assigned. In the first step, the change in the weighted average expression for each condition is evaluated using the gene scores as weights. The obtained average values are called condition scores. Only conditions with a score greater then a threshold t^C are retained. In the second step, the change in the weighted average expression for the retained conditions is evaluated for each gene using the condition scores as weights. These weighted average values are called gene scores. Only genes with a score greater then a threshold t^G are retained. These two steps are repeated until the set of genes and the set of conditions do not change anymore, i.e., a bicluster is identified. Different random initialization and different score thresholds usually result in different biclusters. The algorithm we propose is called AID-ISA and it is summarized in Algorithm 2. It describes the procedure for discovering a single bicluster. In ISA algorithm multiple biclusters are discovered by changing the score thresholds (t^G and t^C) as well as the random seed, and A^G and A^C are built from A by normalizing it in order to have zero mean and unit variance with respect to genes and conditions, respectively. Lines 6-7 and 9-10 refer to the standard ISA implementation. s_i^G (s_j^C) is the weighted score of the ith gene (jth sample); σ^G (σ^C) is the standard deviation of $A^G_{I,\cdot}$ ($A^C_{\cdot,J}$).

AID-ISA calls the AID procedure before the score evaluation steps (lines 5 and 8). In this way, scores are computed over more homogeneous groups. Let us emphasize that AID is performed immediately after the random initialization. This allows the biclustering algorithm to benefit of starting from a set of more homogeneous genes. Like the ISA algorithm, AID-ISA terminates if the processed bicluster does not change anymore. However, since an object may be added and removed alternately by the AID procedure and the standard biclustering step, the algorithm can potentially enter in a infinite loop. To avoid this possibility the bicluster evaluation is stopped if the number of iterations is larger than N (line 11).

4 Experiments and results

In this section we provide experimental evidences of the extra value resulting from additional knowledge injection. We describe the data sources used in the experimental study and how to build the distance matrices from additional information; then, we report the results of a comparative analysis performed over several experimental settings on three gene expression data sets; finally,

Algorithm 2 AID-ISA: Additional Information-Driven ISA

Input: A^G, A^C, m, n, D^G, D^C, t^G, t^C, δ_r, δ_e, N
Output: $B_h = (I, J)$

Initialize: assign to I a random sub set of $\{1, \ldots, m\}$, $s^G = \mathbf{1}^{1\times m}$, $J = \varnothing$, $I' = \varnothing$, $J' = \varnothing$, $n = 0$
repeat
  $n \leftarrow n + 1$
  $I' \leftarrow I$; $J' \leftarrow J$
  AID(I', m, D^G, δ_r, δ_e)
  $s^C \leftarrow s^G_I \times A^G_{I,\cdot}$
  $J' \leftarrow J' \cup \{j' \in \{1, \ldots, n\} | s^C_{j'} \geq t^C \sigma^C\}$
  AID(J', n, D^C, δ_r, δ_e)
  $s^G \leftarrow s^C_J \times (A^C_{\cdot,J})^T$
  $I' \leftarrow I' \cup \{i' \in \{1, \ldots, m\} | s^G_{i'} \geq t^G \sigma^G\}$
until $(I = I' \wedge J = J') \vee (n > N)$

we provide some insights on the ability of AID-ISA in discovering biologically significant biclusters.

4.1 Extraction of the additional information

In this work we assume that features are binary valued vectors. We represent each set of features by a Boolean matrix M. Given an object (gene/condition) p and a feature f, we set M_{pf} to `true` if p has the characteristic described by f, `false` otherwise. In formula:

$$M_{pf} = \begin{cases} \texttt{true} \text{ if } p \text{ has feature } f, \\ \texttt{false} \text{ otherwise.} \end{cases} \quad (1)$$

By leveraging features we compute the distance matrices, D^G and D^C. Specifically, the distance of two objects (p, q) is evaluated using the Tanimoto distance [34]:

$$T_d(p, q) = -log_2 T_s(p, q),$$

where T_s is the Tanimoto similarity, defined as:

$$T_s(p, q) = \frac{\sum_f \mathbf{1}(M_{pf} \wedge M_{qf})}{\sum_f \mathbf{1}(M_{pf} \vee M_{qf})},$$

where $\mathbf{1}(b)$ assumes the value 1 if b is equal to `true`, 0 otherwise. Tanimoto similarity computes the ratio of the number of features characterizing p *and* q, and the number of features characterizing p *or* q.

We notice that the Tanimoto distance violates the triangle inequality and thus it is not a metric. However, in the biological context semi-metrics usually perform better than full-fledged metrics. In fact, the triangle inequality happens to be ill suited to model a frequent situation where two genes involved in no common activities share a common function with a third gene.

Since we aim at measuring to what extent biclusters represent putative transcriptional modules or biological pathways, we choose additional knowledge sources that are coherent with this goal.

For the gene dimension, the Restructured Gene Ontology (RGO) [45] fulfills this purpose suitably. Briefly, the RGO is a reorganization of the Gene Ontology (GO) [2] composed of three ontologies. Each concept is represented by a RGO node, i.e., a group of GO terms that refer to the same regulative activity, and relationships among them are represented by edges. The RGO ontologies are linked by means of cross-ontology edges. These edges connect nodes using a lexical similarity between biological descriptions associated to nodes. Knowledge about genes is codified through their annotations, i.e., associations among RGO nodes and genes. In the RGO, two kinds of annotations are represented: *original* annotations, that are those already present in the GO, and *inferred* annotations, that are derived by following cross-ontologies edges.

We used three sets of features for describing each gene g. Each set refers to annotations that belong to one of the RGO sub-ontologies: RGO Biological Process (BP), RGO Cellular Component (CC), and RGO Molecular Function (MF). These sets identify genes participating to the same biological activities or take into consideration physical closeness of genes. We identify a feature for each RGO node n, instantiating formula (1) as:

$$M_{gn} = \begin{cases} \texttt{true} \text{ if } g \text{ is annotated over } n \\ \texttt{false} \text{ otherwise} \end{cases}$$

To describe each sample s in the sample dimension one can choose features referring to any kind of additional information, coherent with data set at hand, e.g., experimental settings or stress types. In the first data set we used again RGO annotations, whilst in the third data set we exploited patient and cancer characteristics. Specifically, we identify a feature for each patient and cancer characteristic c, instantiating formula (1) as:

$$M_{sc} = \begin{cases} \texttt{true} \text{ if } s \text{ has characteristic } c \\ \texttt{false} \text{ otherwise} \end{cases}$$

4.2 Experimental settings

Data sets To test algorithm performances we used three microarray data sets: two *Saccharomyces cerevisiae* panels that include microarray experiments corresponding to several stress conditions and to gene mutations, and one human panel that contains microarrays of colon-rectal cancer patients.

The first data set, Hughes's panel, consists in 300 microarray experiments corresponding to gene mutations and to treatments with several compounds [18]. We selected 6514 genes having less than 30 missing values and 276 samples corresponding to deletion mutants. Having chosen experimental conditions that explicitly refer to (mutated) genes only has one important side effect: the same features and the same evaluation metric can be used on both the gene and the sample dimensions.

The second data set, Gasch's panel, is composed by 156 microarray experiments that investigate yeast responses to several stresses [14]. We selected 5708 genes having less than 15 missing values. This data set has been chosen in order

Table 1: Population characteristics for the Jorissen's panel.

Characteristic	**Category**	**Patient (n = 60)**
Age	≤ 70	36
	> 70	24
Sex	Female	34
	Male	26
Smoking Status	Non-smokers	56
	Smokers	4
Alcohol Consumption	No	56
	Yes	4
Cancer in family	No	3
	Yes	57
TNM	Low (Stage I + II)	24
	High (Stage III + IV)	36
Metastases	No	24
	Yes	36
Grade	Low	27
	High	33
Localization	Colon and Rectosigmoideum	50
	Rectum	10
Chemotherapy	Yes	22
	No	38
Surgical	Yes	31
	No	29
Radiation	Yes	10
	No	50

to assess the performances of AID-ISA when *a priori* information is available only for one dimension.

The third data set, Jorissen's panel, is composed by 553 microarray experiments on a cohort of patients affected by a primary colon-rectal cancer [22]. We selected 7279 genes having strong profile variations and 60 samples. 30 samples are associated to patients with cancer stage A, i.e., limited invasion; 30 samples correspond to patients with cancer stage D, i.e., widespread cancer with the possibility of metastasis. Each sample is supplied with a set of patient lifestyle and clinical cancer characteristics such as age, gender, tobacco and alcohol consumption, family history of cancer, localization of primary site, tumor grade, and stage of the malignancy. Eventually, information concerning cancer therapies are reported. Table 1 reports the distribution of patient and cancer characteristics. Let us note that characteristics are equally distributed among

the cohort of patients. This data set has been chosen in order to assess the performances of AID-ISA when different *a priori* information is available for gene and sample dimension.

Parameter settings Thresholds were set as suggested in [19, 10]. Specifically, threshold t^G ranged from 1.8 to 4 (in steps of 0.1), and t^C was fixed to 2. These values were used in both ISA and AID-ISA. In AID we set δ_r and δ_e to 2 and 0.5, respectively. These values allow the deletion of very far objects and the addition of very close objects. In this way, on the one hand, only completely unrelated objects are discarded and, on the other hand, the noise is kept under control and biclusters cannot grow arbitrarily. The value of N in AID-ISA was set to 100. For each experiment 20 runs were performed and results were averaged.

Bicluster evaluation To assess the quality of the obtained biclusters in the *S. cerevisiae* data sets we used the Biological Homogeneity Index (BHI) [11]. The BHI measures whether, on average, genes belonging to the same cluster also belong to the same functional class. It is evaluated as:

$$BHI(C) = \frac{1}{h}\sum_{i=1}^{h}\frac{1}{n_i(n_i-1)}\sum_{p,q\in C_i, p\neq q}\mathbf{1}\left(\Phi(p)\cap\Phi(q)\neq\varnothing\right),$$

where Φ is a function mapping each gene $g \in G$ to a subset of the functional classes $F = \{\mathtt{f}_1, \ldots, \mathtt{f}_k\}$ describing its activity (specifically, $\Phi(g) = \{\mathtt{f}_i \in F \mid g \text{ is annotated over } \mathtt{f}_i\}$), and n_i is the number of functionally annotated genes in C_i, i.e., $n_i = |\{g \in C_i | \Phi(g) \neq \varnothing\}|$. BHI ranges between 0 and 1; high BHI values are better.

GO or RGO annotations may be chosen as functional classes. However, to ensure a fairer evaluation, we chose to use different information, that is the gene mutant phenotypes collected by the SGD project [35].

Let us note that BHI has been designed to evaluate clusters and cannot be applied to biclustering algorithms directly. However, since in the Hughes's panel conditions explicitly refer to genes, we still applied this metric to the biclustering results by evaluating each dimension separately. Since functional classes are not available for human data, the BHI measure cannot be used to evaluate biclusters obtained on the Jorissen's panel and a manual validation was performed (see Section 4.4).

4.3 Performance evaluation

We performed two kinds of experiments: in the first one we show that AID-ISA outperforms ISA in discover functionally enriched groups. In the second experiment we rule out the hypothesis that AID process may yield to comparable results when used as a mere data post-processing.

Hughes's panel Table 2 shows the results obtained by the ISA algorithm. ISA obtains good results when the BHI is evaluated on gene clusters. This outcome is not surprising: ISA is considered one of the best approaches for identifying functional enriched biclusters [32]. Nevertheless, BHI value dramatically decreases when clusters of samples are examined.

Table 2: BHI values obtained by the ISA algorithm on Hughes's panel.

	Average BHI	**Standard Deviation**
gene	0.930	0.005
sample	0.089	0.044

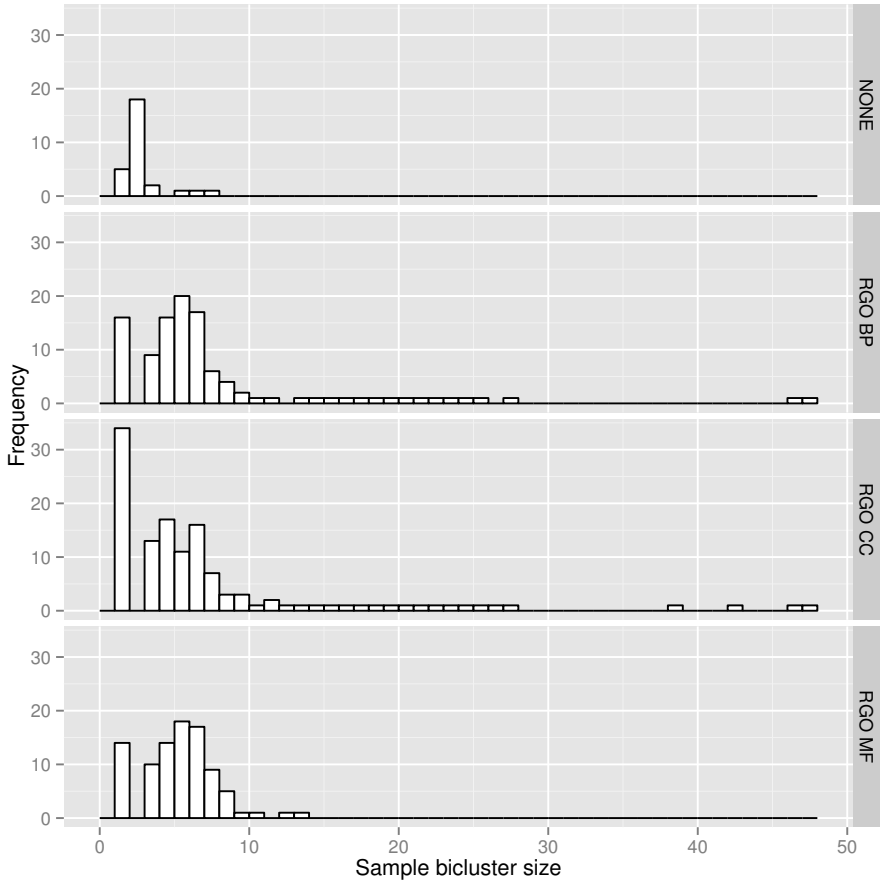

Figure 1: Size of biclusters found by ISA and AID-ISA algorithms on Hughes's panel, sample dimension. The histograms show, for each set of features, the distribution of the sample size. The plot referred as *"None"* shows ISA results.

Poor performances over the sample dimension are mainly due to the size of the discovered biclusters (see Figure 1, top panel). Indeed, ISA creates many biclusters grouping few samples (we will refer to this issue as the "sample-dimension bicluster-size problem"). For instance, 17 out of the 21 identified biclusters group only two samples. Thus, ISA is able to identify biclusters of functionally enriched genes, but only in a small subset of samples.

Columns 1 and 2 of Table 3 report, for each set of features, its size and the percentage of genes having at least one annotation. We recall that only

genes/samples having at least one associated information are included in the distance matrices. Columns 3 and 4 of Table 3 show the results obtained by AID-ISA. AID-ISA increases BHI values with respect to those obtained by ISA, especially on, but not limited to, the sample dimension. Moreover, the size of biclusters on the sample dimension is better spread with respect to that obtained by the ISA algorithm (see Figure 1). It can be argued that this property makes biclusters more meaningful from a biological point of view, as we will show in Section 4.4.

Table 3: BHI values obtained by the AID-ISA algorithm on Hughes's panel when additional information is extracted from the RGO.

Feature set		**Features**		**BHI**	
		Number of features	**% annot. genes**	**Average**	**Standard Deviation**
RGO-BP	gene	15,589	83.8%	0.940	0.005
	sample		70.0%	0.838	0.090
RGO-CC	gene	2,918	83.8%	0.935	0.002
	sample		70.0%	0.673	0.060
RGO-MF	gene	9,149	83.8%	0.939	0.006
	sample		70.0%	0.849	0.072

Table 4 shows BHI values obtained by using AID as a post-processing step. Let us note that BHI values evaluated on gene clusters are always larger than those obtained by ISA. However, a simple post-processing cannot avoid the "sample-dimension bicluster-size problem". On the contrary, it seems that a post-processing driven approach leads to even worse performances. This is due to the small size of the original biclusters. Indeed, when the AID process is applied on few genes the result is unpredictable and usually meaningless.

Table 4: BHI values obtained by applying AID on Hughes's panel in different ways.

Feature set		**AID post-processing**		**AID-ISA**	
		Average BHI	**Standard Deviation**	**Average BHI**	**Standard Deviation**
RGO-BP	gene	0.939	0.004	0.940	0.005
	sample	0.020	0.009	0.838	0.090
RGO-CC	gene	0.938	0.003	0.935	0.002
	sample	0.017	0.009	0.673	0.060
RGO-MF	gene	0.938	0.002	0.939	0.006
	sample	0.018	0.009	0.849	0.072

Gasch's panel In this data set additional information is available only for genes and both the AID process and the cluster evaluation have been performed only on the gene dimension.

Table 5: BHI values obtained by ISA and AID-ISA algorithm when features are extracted from the RGO on Gasch's panel. Only gene dimension is considered. The row referred as *"None"* refers to the ISA results.

	Features		**BHI**	
Feature set	**Number of features**	**% of annotated genes**	**Average**	**Standard Deviation**
none	-	-	0.746	0.017
RGO-BP	15,589	88.9%	0.978	0.014
RGO-CC	2,918	88.9%	0.949	0.025
RGO-MF	9,149	88.9%	0.952	0.027

Table 5 reports, for each set of features, its size and the percentage of genes having at least one annotation (Columns 1 and 2) and the BHI values we obtained when both the ISA (Row 1) and the AID-ISA (Rows 2-4) algorithms are applied (Columns 3 and 4). AID-ISA always returns BHI values larger than those obtained by the ISA approach.

Table 6: BHI values obtained by applying AID in different ways on Gasch's panel. Only gene dimension is considered.

	AID post-processing		**AID-ISA**	
Feature set	**Average BHI**	**Standard Deviation**	**Average BHI**	**Standard Deviation**
RGO-BP	0.950	0.001	0.978	0.014
RGO-CC	0.955	0.001	0.949	0.025
RGO-MF	0.951	0.001	0.952	0.027

Table 6 shows the different performances obtained by using AID as a post-processing step or inside the ISA algorithm. The performances obtained by leveraging the AID process are again better than those obtained by ISA (Table 5, Row 1), while the post-processing by means of the AID process slightly outperforms AID-ISA only when features describing the cellular components are used.

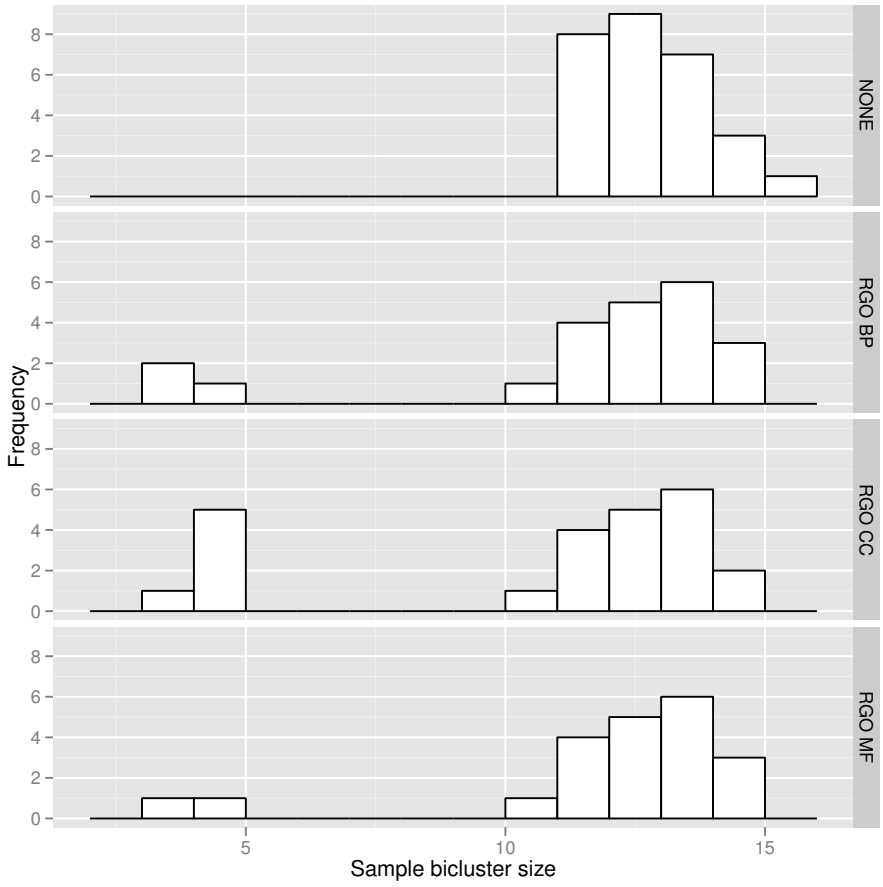

Figure 2: Size of biclusters found by ISA and AID-ISA algorithms on Gasch's panel, s

As a last remark about this data set, let us underline that, even though no constraint has been set on the sample dimension, the size of biclusters on this dimension is well spread (see Figure 2). Notwithstanding, nothing hinders to choose features referring to any kind of additional information, e.g., stress types. The investigation of multiple features will be discussed in the following section.

Jorissen's panel In this data set two sources of information are available: RGO annotations for the gene dimension; patient lifestyle and clinical cancer characteristics for the sample dimension. Since we cannot compute BHI values, we proceed with a manual evaluation of biclusters obtained by the ISA and the AID-ISA algorithms. The coherence of the obtained results will be discussed in Section 4.4. Let us only remark that the size of biclusters on genes dimension is again well spread (see Figure 3).

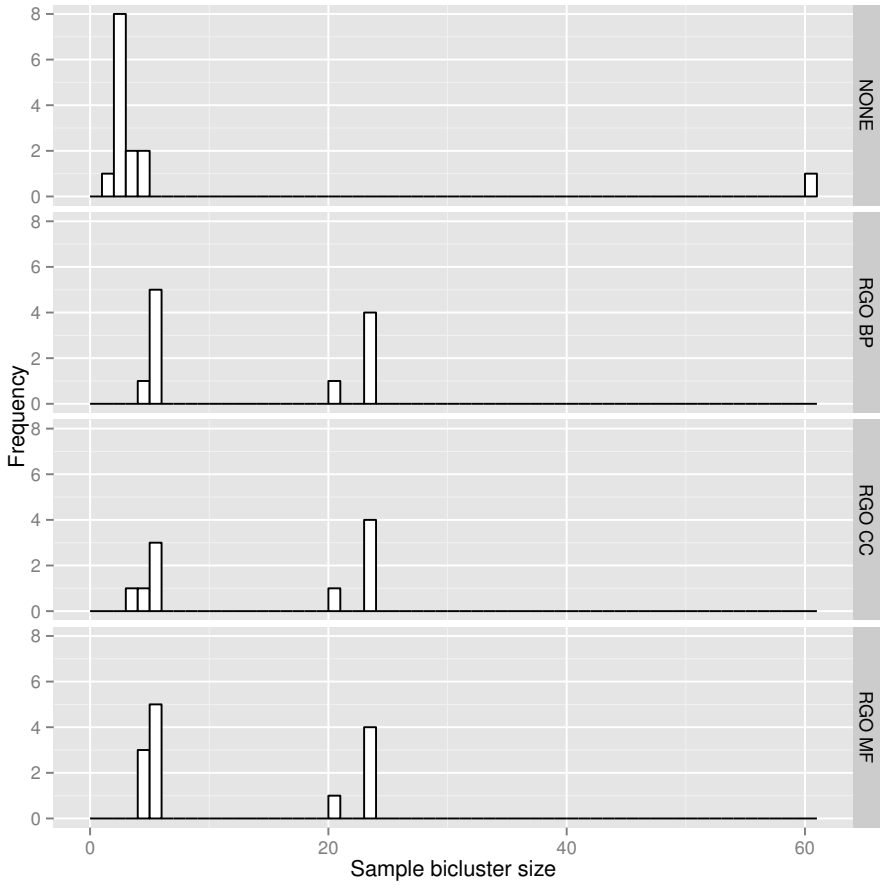

Figure 3: Size of biclusters found by ISA and AID-ISA algorithms on Jorissen's panel, sample dimension. Histograms show, for each set of features, the distribution of the sample size. The plot referred as *"None"* shows ISA results.

4.4 Evaluation of bicluster biological coherence

We performed an in-depth analysis to assess the biclusters coherence and quality from a biological point-of-view. In the following we focus on biclusters returned by AID-ISA using RGO BP features. A bicluster has been chosen for the Hughes's and the Gasch's data sets, whilst a deeper analysis has been performed on the Jorissen's data set. In the selected bicluster, we will refer to the set of genes with symbol I and to the set of samples with symbol J.

Hughes's panel We extracted 83 biclusters from a randomly chosen run of the AID-ISA algorithm. They show a BHI value of 0.935 for the gene dimension and a BHI value of 0.831 for the sample dimension. Within this result set, to allow a manual evaluation, we focused on its smallest member. It groups 39 genes and 3 different experimental conditions (the mutant genes HST3, TUP1, and SSN6).

To asses the quality of the association between genes and samples, we analyzed the biological functions that both groups perform. Mutant genes belonging to J are involved in metabolic activities. In particular, the SSN6-TUP1 protein complex is involved in the metabolism of carbon sources [1]. HST3 is involved in short-chain fatty acid metabolism [41] and calorie restrictions may interfere with HST3 activity [24]. A functional enrichment of genes in I performed using the FunSpec web application [33] and the MIPS Functional Classification [28] reveals that they are involved in "sugar transport", "metabolism", and "metabolism of energy reserves" (p-values < 0.001). Summarizing, both genes in I and in J are involved in metabolic processes, thus confirming the faithfulness of the obtained bicluster.

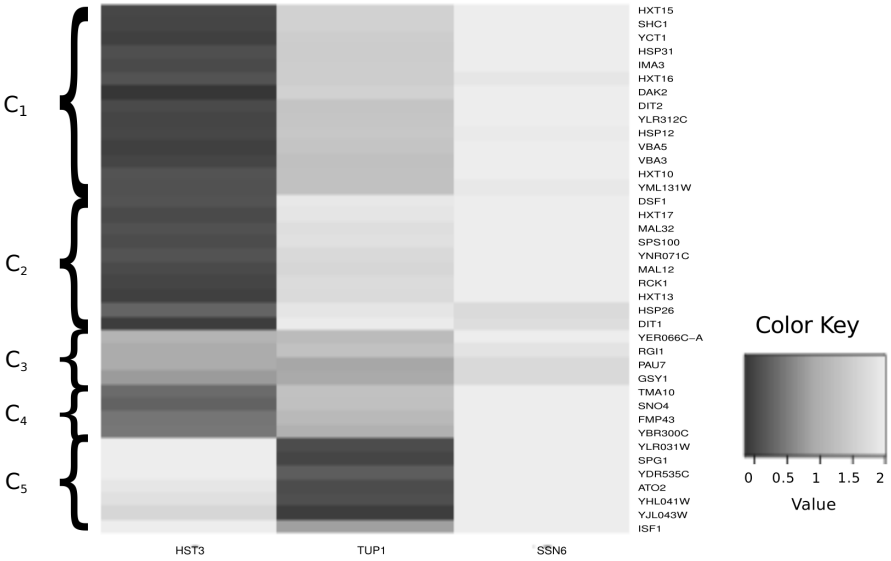

Figure 4: Bicluster heat map. Rows represent genes belonging to I, and columns represent genes (mutant conditions) belonging to J. Each cell (i, j) is filled based on the expression level of gene i in condition j. Labels on the left show a possible separation on the gene dimension based on expression levels.

Hereafter, we point out evidences suggesting that the genes in the selected bicluster form a transcriptional module: i.e., they are co-expressed and bound by the same transcription factors. Figure 4 shows the heat map of the selected bicluster. In contrast to what one would expect, the genes do not appear to have a very close expression profile. In fact, as shown by the labels on the left, one can recognize five different groups of co-expressed genes. The following argument shows that they participate to the same process nonetheless, showing that AID-ISA allows the discovery of transcriptional modules that could not be recovered using expression profiles alone. First, we used the Yeast Promoter Atlas [6] to obtain the list of transcription factors binding genes in I. Among the found transcription factors are: SPT15 that binds 11 of the genes in I, MSN2 that binds 8 genes, and NRG1 that binds 5 genes. Interestingly, the

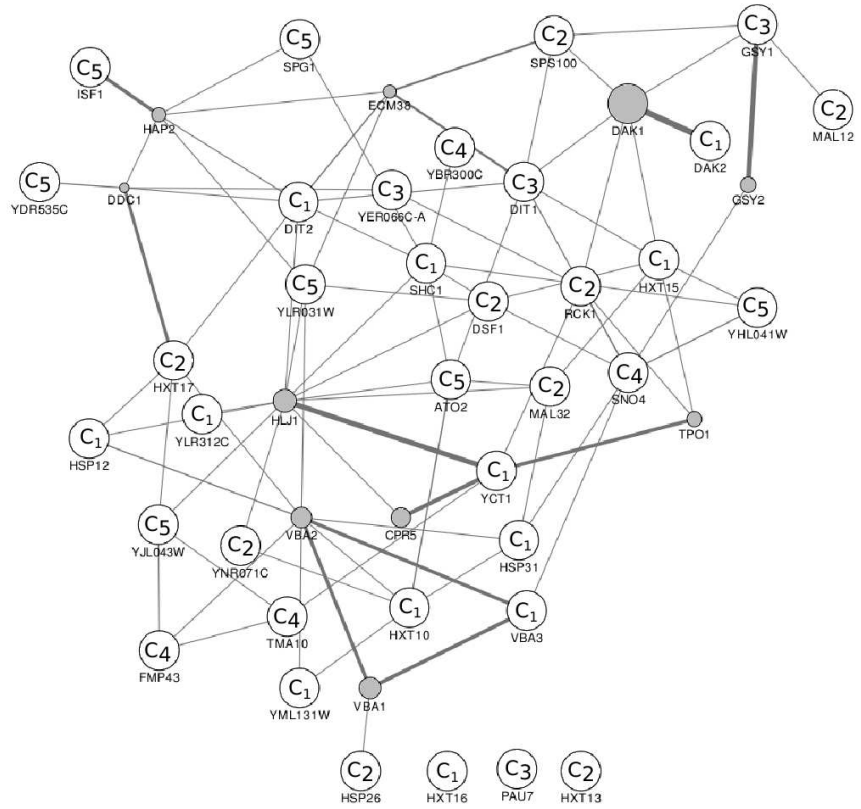

Figure 5: Interactions among genes biclustered together (gene dimension). Genes in to I are labeled according to clustering reported in Figure 4. Edges represent genetic interactions. The network is built and visualized using the GeneMANIA Cytoscape plugin.

genes regulated by the these transcription factors are distributed on the five co-expressed groups we mentioned before. Then, despite being in different co-expression groups, there exist transcriptional interactions among them. The second piece of evidence we provide confirms this finding. Figure 5 shows the genetic interactions among genes belonging to I as created by GeneMANIA Cytoscape plugin [29, 36]. GeneMANIA leverages data collected from several primary studies and from BioGRID database and reports information about gene interactions, setting edges among genes if they are functionally associated. It is important to point out that gene interactions are often functionally relevant since they impact on gene expression when organisms are subject to several stress conditions.

Gasch's panel To describe the obtained results we perform an analysis close to that described for the Hughes's panel. We comment a randomly chosen run of the AID-ISA algorithm. It extracts 16 biclusters showing a BHI value of 0.98 for the gene dimension. Within this result set, we focused on its smallest member, which groups 34 genes and 13 experimental conditions.

To assess the quality of the extracted bicluster, we analyzed the biological functions that genes in I perform. We found that *i)* they are strictly related to the experimental conditions in J, and that *ii)* they form a transcriptional module.

Experimental conditions in J correspond to cell response to amino acid starvation. Carbon and nitrogen starvation dramatically affects gene expression programs, and leads to growth arrest and entrance into a stationary phase, causing also the interruption of protein synthesis [14, 46]. A number of genes induced during growth to stationary phase contain elements responsive to cAMP, a well-know transcription factor, such as the CTT7-STRE element [46]. Interestingly, in addition to nutrient starvation, the CTT7-STRE element can also be activated by other environmental stresses such as heat shock and osmotic stress [46].

A functional enrichment of genes in I performed using again the FunSpec web application and the MIPS Functional Classification reveals that they are involved (p-values < 0.005) in "oxidative stress response", in "heat shock response", and, notably in "biosynthesis of glutamate" (a cellular process moderately triggered during amino acid starvation condition [44]).

Summarizing, both genes in I and in J are involved in nutrition depletion, thus confirming the biological coherence of the obtained bicluster.

To check whether genes in I are bound by the same transcription factors, we used the Yeast Promoter Atlas. Among the found transcription factors are: MSN2 that binds 14 genes, UME6 that binds 6 genes, and NRG1 that binds 5 genes. All these transcription factors are related to a cell starvation phase: MNS2 is activated in several stress condition, including amino acid starvation [27]; UME6 is responsible of metabolic responses to nutritional cues [47]; and NRG1 negatively regulates a number of processes [23].

Jorissen's panel Biclusters resulting by applying ISA and AID-ISA to the Jorissen's data set show important overlaps. Then, as suggested in [20], we cleaned up the results by selecting only those biclusters having an overlap smaller than the 75%. We ended up with 2 biclusters for the AID-ISA processing, and 8 biclusters for the ISA one. We selected one bicluster from the AID-ISA result, grouping 361 genes and 23 samples, and 2 biclusters from the ISA results, grouping 165 and 232 genes respectively. Both ISA biclusters include only 2 samples. We chose these biclusters because they group genes with the most similar size. Let us note that all the biclusters returned by ISA involve at most 5 samples.

To assess the quality of the extracted bicluster, we analyzed the biological functions that genes in I perform, showing that they are closely related to the clinical characteristics that are described by samples in J.

We used the Ingenuity Pathway Analysis software (IPA 7.0, Ingenuity System[1]) to show the biological coherence of gene clusters. Specifically, we used the IPA tool to functionally annotate genes according to biological processes and canonical pathways, and to identify genes potentially associated to cancer and other diseases. We manually evaluated the coherence of sample clusters by analyzing patient and cancer characteristics.

[1] `http://www.ingenuity.com/`

First we analyze the AID-ISA result. 22 out of 23 patients in J have been diagnosed with a severe tumor, showing also widespread metastases. According to the IPA analysis, the most statistically significant canonical pathways enriched by genes in I (p-value < 0.001) are "mammalian target of rapamycin *(mTOR)* signaling", "regulation of eukaryotic initiation factor 4 *(eIF4)* and p70 S6 kinase signaling", and "eukaryotic initiation factor 2 *(eIF2)* signaling".

Let us briefly comment these findings. mTOR is a regulator of protein translations and cell metabolism, allowing cells to grow and proliferate [49]. Moreover, a recent study shows that mTOR plays a crucial role in regulating cancer cell migration, and cancer metastasis [50]. eIF4 is often overexpressed in human cancers, and, in experimental models, it has been related to disease progression, cellular transformation, tumorigenesis, and metastatic progression [15] eIF2 has also been showed to be connected to cancer development [39].

Summarizing, the bicluster identified by AID-ISA associates a subset of patients experiencing widespread metastases with a set of pathways strictly connected to a metastatic cancer behavior.

Biclusters obtained by the ISA algorithms, according to the IPA analysis, enrich canonical pathways related to cancer progression and invasion, e.g., "eIF2 signaling", and "DNA damage checkpoint regulation in cell cycle" (p-values < 0.001). However, in this case it is difficult to discover a significant pattern given that only 2 patients are returned, and, as a consequence, also to find an association between these patients and cancer progression and invasion. This is yet another evidence supporting the claim that AID-ISA provides biologically relevant biclusters, solving the "sample-dimension bicluster-size problem" that affects ISA.

5 Conclusions

In this paper we proposed a new biclustering approach that embeds a refinement process leveraging additional information into the well-known Iterative Signature Algorithm. In detail, we described: *i)* a general algorithm, AID, that exploits additional knowledge; *ii)* a modified version of ISA, AID-ISA, that implements our proposal, and *iii)* an approach to extract functional information from the Restructured Gene Ontology and clinical data.

Let us underline that the definition of our refinement module is very general: it only requires the availability of some additional features and a distance metric defined on them. Thus, it can be used with any source of information and it can be embedded into other iterative biclustering schemes. We ended up with ISA because it has been shown as being the most effective among competitors [32].

A comprehensive set of experiments, performed on baker's yeast and human gene expression data sets, showed that the biclusters extracted by AID-ISA provide more reliable and more complete biological insights than those returned by ISA. The identification of coherent gene groups is a key task in regulatory genomics. For instance, biclusters are used to predict genes functions by the assumption of a "guilt-by-association" heuristic [48], i.e., a gene is predicted to have the same functions of genes clustered with it. When biclustering process uses only the information derived by gene expression profiles, the application of this heuristic may lead to controversial outcomes. In fact, other types of associations, such as cis-motif co-occurrence, are more strongly tied to gene functions than co-expression. The adoption of a grouping criterion based on

both expression analysis and functional annotations enhances the discovery of groups of genes that shares the same functions.

Nowadays, Next Generation Sequencing technologies are replacing microarrays for transcriptome expression profiling. To exploit these new and precious pieces of information we will adapt the AID approach to the analysis of transcript expression levels obtained from these experiments.

AID-ISA is available as source code[2]. As future work, we will set up a Web application to support various standard and personalized bicluster discovery scenarios with the automatic retrieval of up-to-date GO and RGO graphs and other sources of information.

Acknowledgments

We would like to thank prof. Raffaele A. Calogero for providing the access to the Ingenuity Pathway Analysis tool, and Dr. Roberto Esposito for commenting and proof-reading the manuscript.

References

[1] L. Alberghina, G. Mavelli, G. Drovandi, P. Palumbo, S. Pessina, F. Tripodi, P. Coccetti, and M. Vanoni. Cell growth and cell cycle in Saccharomyces cerevisiae: Basic regulatory design and protein-protein interaction network. *Biotechnology advances*, 2011.

[2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, 2000.

[3] C. Becquet, S. Blachon, B. Jeudy, J-F. Boulicaut, and O. Gandrillon. Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. *Genome Biology*, 3(12), 2002.

[4] J. Besson, C. Robardet, and S. Boulicaut, J-F.and Rome. Constraint-based concept mining and its application to microarray data analysis. *Intell. Data Anal.*, 9(1):59–82, 2005.

[5] M. Brameier and C. Wiuf. Co-clustering and visualization of gene expression data and gene ontology terms for Saccharomyces cerevisiae using self-organizing maps. *Journal of biomedical informatics*, 40(2):160–173, 2007.

[6] D.T.H. Chang, C.Y. Huang, C.Y. Wu, and W.S. Wu. YPA: an integrated repository of promoter features in Saccharomyces cerevisiae. *Nucleic acids research*, 39(suppl 1):D647–D652, 2011.

[2] `http://compbio.di.unito.it/tools/AID-ISA/`

[7] Y. Cheng and G.M. Church. Biclustering of expression data. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology; ISMB*, volume 8, page 93, 2000.

[8] Mark H Chin, Mike J Mason, Wei Xie, Stefano Volinia, Mike Singer, Cory Peterson, Gayane Ambartsumyan, Otaren Aimiuwu, Laura Richter, Jin Zhang, et al. Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. *Cell stem cell*, 5(1):111–123, 2009.

[9] H. Cho, I. S. Dhillon, Y. Guan, and S. Sra. Minimum Sum-Squared Residue Co-Clustering of Gene Expression Data. In *Proceedings SIAM SDM 2004*, Lake Buena Vista, FL, USA, 2004.

[10] G. Csárdi, Z. Kutalik, and S. Bergmann. Modular analysis of gene expression data with R. *Bioinformatics*, 26(10):1376, 2010.

[11] S. Datta and S. Datta. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC bioinformatics*, 7(1):397, 2006.

[12] Tomislav Domazet-Lošo and Diethard Tautz. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, 468(7325):815–818, 2010.

[13] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863, 1998.

[14] A.P. Gasch, P.T. Spellman, C.M. Kao, O. Carmel-Harel, M.B. Eisen, G. Storz, D. Botstein, and P.O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell*, 11(12):4241, 2000.

[15] Jeremy R Graff, Bruce W Konicek, Julia H Carter, and Eric G Marcusson. Targeting the eukaryotic translation initiation factor 4E for cancer therapy. *Cancer research*, 68(3):631–634, 2008.

[16] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18(suppl 1):S145–S154, 2002.

[17] Norman Huang, Parantu K Shah, and Cheng Li. Lessons from a decade of integrating cancer copy number alterations with gene expression profiles. *Briefings in bioinformatics*, 13(3):305–316, 2012.

[18] T.R. Hughes, M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H. Dai, Y.D. He, Y.D. Matthew, J. Kidd, A.M. King, M.R. Meyer, D. Slade, P.Y. Lum, S.B. Stepaniants, D.D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S.H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, 2000.

[19] J. Ihmels, S. Bergmann, and N. Barkai. Defining transcription modules using large-scale gene expression data. *Bioinformatics*, 20(13):1993, 2004.

[20] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. Revealing modular organization in the yeast transcriptional network. *Nature genetics*, 31(4):370–378, 2002.

[21] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 16(11):1370–1386, 2004.

[22] RN. Jorissen, P. Gibbs, M. Christie, S. Prakash, L. Lipton, J. Desai, D. Kerr, LA. Aaltonen, D. Arango, M. Kruhffer, TF. Orntoft, CL. Andersen, M. Gruidl, VP. Kamath, S. Eschrich, TJ. Yeatman, and OM. Sieber. Metastasis-Associated Gene Expression Changes Predict Poor Outcomes in Patients with Dukes Stage B and C Colorectal Cancer. *Clinical Cancer Research*, 15(24):7642–7651, 2009.

[23] Sergei Kuchin, Valmik K Vyas, and Marian Carlson. Snf1 protein kinase and the repressors Nrg1 and Nrg2 regulate FLO11, haploid invasive growth, and diploid pseudohyphal differentiation. *Molecular and cellular biology*, 22(12):3994–4000, 2002.

[24] S.P. Lu and S.J. Lin. Regulation of yeast sirtuins by NAD+ metabolism and calorie restriction. *Biochimica et Biophysica Acta (BBA)-Proteins & Proteomics*, 1804(8):1567–1575, 2010.

[25] S. C. Madeira and A. L. Oliveira. Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45, 2004.

[26] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, Gustavo Stolovitzky, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 2012.

[27] Oliver Medvedik, Dudley W Lamming, Keyman D Kim, and David A Sinclair. MSN2 and MSN4 link calorie restriction and TOR to sirtuin-mediated lifespan extension in Saccharomyces cerevisiae. *PLoS biology*, 5(10):e261, 2007.

[28] HW Mewes, D Frishman, U. Güldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Münsterkötter, S. Rudd, and B. Weil. MIPS: a database for genomes and protein sequences. *Nucleic acids research*, 30(1):31, 2002.

[29] J. Montojo, K. Zuberi, H. Rodriguez, F. Kazi, G. Wright, SL Donaldson, Q. Morris, and GD Bader. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics*, 26(22):2927–2928, 2010.

[30] R. G. Pensa and J-F. Boulicaut. Constrained Co-clustering of Gene Expression Data. In *Proceedings SIAM SDM 2008*, pages 25–36, Atlanta, GA, USA, 2008.

[31] R.G. Pensa, J.F. Boulicaut, F. Cordero, and M. Atzori. Co-clustering numerical data under user-defined constraints. *Statistical Analysis and Data Mining*, 3(1):38–55, 2010.

[32] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122, 2006.

[33] M.D. Robinson, J. Grigull, N. Mohammad, and T.R. Hughes. FunSpec: a web-based cluster interpreter for yeast. *BMC bioinformatics*, 3(1):35, 2002.

[34] D. J. Rogers and T. T. Tanimoto. A Computer Program for Classifying Plants. *Science*, 132:1115–1118, 1960.

[35] Saccharomyces Genome Database. Saccharomyces Phenotype Terms. `http://www.yeastgenome.org/cache/PhenotypeTree.html`, 2012.

[36] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[37] X. Shi, W. Fan, and P.S. Yu. Efficient Semi-supervised Spectral Co-clustering with Constraints. In *Proceedings of ICDM 2010*, pages 1043–1048, Sydney, Australia, 2010.

[38] Robert H Shoemaker. The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer*, 6(10):813–823, 2006.

[39] Deborah Silvera, Silvia C Formenti, and Robert J Schneider. Translational control in cancer. *Nature Reviews Cancer*, 10(4):254–266, 2010.

[40] Y. Song, S. Pan, S. Liu, F. Wei, M.X. Zhou, and W. Qian. Constrained Co-clustering for Textual Documents. In *Proceedings of AAAI 2010*, Atlanta, Georgia, USA, 2010. AAAI Press.

[41] V.J. Starai, H. Takahashi, J.D. Boeke, and J.C. Escalante-Semerena. Short-chain fatty acid activation by acyl-coenzyme A synthetases requires SIR2 protein function in Salmonella enterica and Saccharomyces cerevisiae. *Genetics*, 163(2):545–555, 2003.

[42] D. Steinhauser, B.H. Junker, A. Luedemann, J. Selbig, and J. Kopka. Hypothesis-driven approach to predict transcriptional units from gene expression data. *Bioinformatics*, 20(12):1928–1939, 2004.

[43] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18(suppl 1):S136–S144, 2002.

[44] Lourdes Valenzuela, Paola Ballario, Cristina Aranda, Patrizia Filetici, and Alicia González. Regulation of expression of GLT1, the gene encoding glutamate synthase in Saccharomyces cerevisiae. *Journal of bacteriology*, 180(14):3533–3540, 1998.

[45] A. Visconti, R. Esposito, and F. Cordero. Restructuring the Gene Ontology to emphasise regulative pathways and to improve gene similarity queries. *International Journal of Computational Biology and Drug Design*, 4(3):220–238, 2011.

[46] M. Werner-Washburne, EL Braun, ME Crawford, and VM. Peck. Stationary phase in Saccharomyces cerevisiae. *Molecular Microbiology*, 19(6):1159–1166, 1993.

[47] Roy M Williams, Michael Primig, Brian K Washburn, Elizabeth A Winzeler, Michel Bellis, Cyril Sarrauste de Menthière, Ronald W Davis, and Rochelle E Esposito. The Ume6 regulon coordinates metabolic and meiotic gene expression in yeast. *Proceedings of the National Academy of Sciences*, 99(21):13431–13436, 2002.

[48] C.J. Wolfe, I.S. Kohane, and A.J. Butte. Systematic survey reveals general applicability of guilt-by-association within gene coexpression networks. *BMC Bioinformatics*, 6:227, 2005.

[49] Stephan Wullschleger, Robbie Loewith, and Michael N Hall. TOR signaling in growth and metabolism. *Cell*, 124(3):471–484, 2006.

[50] H. Zhou and S. Huang. Role of mTOR signaling in tumor cell motility, invasion and metastasis. *Curr Protein Pept Sci.*, 12(1):30–42, 2011.