Improving the Performance of the Iterative Signature Algorithm for the Identification of Relevant Patterns

A. Freitas[1,2*], V. Afreixo[1,2], M. Pinheiro[3,4], J. L. Oliveira[3,5], G. Moura[6,7] and M. Santos[6,7]

[1]*Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal*

[2]*CIDMA, University of Aveiro, 3810-193 Aveiro, Portugal*

[3]*Department of Electronics, Telecommunications and Informatics, University of Aveiro, 3810-193 Aveiro, Portugal*

[4]*Biocant, Bioinformatics Unit, 3060-197 Cantanhede, Portugal*

[5]*IEETA, University of Aveiro, 3810-193 Aveiro, Portugal*

[6]*Department of Biology, University of Aveiro, 3810-193 Aveiro, Portugal*

[7]*CESAM, University of Aveiro, 3810-193 Aveiro, Portugal*

Received 30 August 2009; revised 11 June 2010; accepted 30 November 2010
DOI:10.1002/sam.10104
Published online 13 January 2011 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: The iterative signature algorithm (ISA) has become very attractive to detect co-regulated genes from microarray data matrices and can be a useful tool for the identification of similar patterns in many other kinds of numerical data matrices. Nevertheless, its algorithmic strategy exhibits some limitations since it is based on statistical behavior of the average and considers averages weighted by scores not necessarily positive. Hence, we propose to take the median instead of the average and to use absolutes scores in ISA's structure. Furthermore, a generalized function is also introduced in the algorithm in order to improve its algorithmic strategy for detecting high value or low value biclusters. The effects of these simple modifications on the performance of the biclustering algorithm are evaluated through an experimental comparative study involving synthetic data sets and real data from the organism *Saccharomyces cerevisiae*. The experimental results show that the proposed variations of ISA outperform the original version in many situations. Absolute scores in ISA are shown to be essential for the correct interpretation of the biclusters found by the algorithm. The median instead of the average turns the biclustering algorithm more resilient to outliers in the data sets. © 2011 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 4: 71–83, 2011

Keywords: biclustering; iterative signature algorithm; DNA; microarray; codon; median; average

1. INTRODUCTION

The increasing number of sequenced genomes and the large amount of complex data emerging from DNA microarray technologies have created new challenges in several scientific domains, namely statistics and computational sciences. An important challenge is the identification of patterns or homogeneous groups. For instance, in studies of gene primary structure features, the detection of similar patterns of codon-pair context, in fully sequenced genomes, can be important to unveil general rules that influence the mRNA decoding fidelity [1–3]. Also, in the analysis of gene expression data resulting from DNA microarray experiments, the identification of genes, with similar expression profiles under the same subset of experimental conditions, is fundamental for the identification of regulatory properties of cellular processes [4].

The potential of clustering methods to reveal biologically meaningful patterns was initially considered by Eisen *et al*. [5], who applied hierarchical clustering to identify

Correspondence to: A. Freitas (adelaide@ua.pt)

functional groups of genes. After that, several clustering methods for gene expression data have been introduced and evaluated [6]. Nevertheless, standard clustering techniques have shown some limitations. For instance, in microarray data sets, these methods do not allow overlapped clusters. Hence, they are not adequate for biological systems where the same gene may be involved in multiple processes and therefore belong to multiple clusters.

To overcome some of these limitations, new approaches of clustering have been proposed in the last years [4,7–9]. These algorithms detect groups considering, simultaneously, the two dimensions of the data matrix and are called biclustering. In Ref. 10, several types of biclusters are discussed. Biclustering is a NP-hard problem [7], and no solution is optimal for finding optimal sets of biclusters. Each algorithm is defined by one particular criterion of biclustering and has its own advantages and disadvantages. Recently, Prelic *et al*. [11] addressed an empirical comparative study of five different biclustering methods. In contrast, we focus our study on only one biclustering algorithm, and we propose modifications in order to improve its performance.

The iterative signature algorithm (ISA) is a biclustering algorithm able to obtain overlapping biclusters. It was originally proposed by Ihmels *et al*. [4,8] to identify transcription modules from microarray experiments. A transcription module consists of a set of co-regulated genes and an associated set of regulating conditions. In Ref. 11, it is shown that ISA provides good results on various synthetic and real data sets.

Let $\mathbf{X}$ be a $n \times m$ matrix of real numbers given by

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} = \left[x_{ij}\right] \qquad (1)$$

where the n rows are denoted by $R_1, R_2, \ldots, R_n$ and the m columns are $C_1, C_2, \ldots, C_m$. Running ISA one time on $\mathbf{X}$ originates as much as one single bicluster. Running several times, multiple biclusters, overlapped or not, can be detected. Each bicluster is expected to be a submatrix of $\mathbf{X}$ whose observations, for each row and each column, have weighted averages that do not belong to specified intervals predefined in terms of two threshold parameters. Basically, ISA starts with a subset of rows (randomly chosen or not) and applies iteratively the signature algorithm introduced in Ref. 4 until two consecutive iterations yield the same set of rows. The signature algorithm is well described by its authors in Refs. 4,8. It is processed in two stages considering the matrix $\mathbf{X}$ normalized by columns (rows) in the first (second) stage. Starting with an initial set of rows and uniform row scores, columns whose averages weighted by the row scores that do not belong to a bounded interval $\mathcal{I}_x$, predefined in terms of a threshold t_x, are selected. In the second stage, and for the columns chosen in the first stage, the algorithm selects all the rows whose averages, weighted by the column scores, exceed a limit a_y predefined in terms of a threshold t_y, that is, that do not belong to the interval $\mathcal{I}_y = \,]-\infty, a_y]$. The column (row) scores are the weighted averages calculated by the column (row) in the immediately previous stage (iteration). These weights are the row (column) scores obtained in the immediately previous iteration (stage) of the algorithm.

If all those weights are assumed to be equal to one, $\mathcal{I}_x$ ($\mathcal{I}_y$) is $[\hat{\mu} - t_x\hat{\sigma}, \hat{\mu} + t_x\hat{\sigma}]$ ($]-\infty, \hat{\mu} + t_y\hat{\sigma}]$), which can represent a confidence interval, with the level of confidence of $(1-\alpha) \times 100\%$, for the mean by column (row), if each column (row) comes from a Gaussian distribution or if the Central Limit Theorem holds. In this case, t_x (t_y) is the quantile of order $1-\alpha/2$ ($1-\alpha$) of the standard Gaussian distribution and $\hat{\mu}$ and $\hat{\sigma}$ are the estimated mean and standard deviation of the sample average. Thus, if unit scores were considered, we could say that the algorithm would search biclusters whose rows and columns belong to critical regions of hypothesis tests for testing the mean defined in terms of z-score statistics. Hence, the use of normalized or non-normalized data in the algorithm would be equivalent.

We have implemented ISA in a software platform named Anaconda [2], which has been created by us as a way of studying codon-pair context biases in fully sequenced genomes [1,3]. Basically, for each sequenced genome, Anaconda imports complete sets of Open Reading Frames from public databases and converts them into codon-pair contingency tables. Each contingency table (64 rows $\times$ 64 columns) is associated to the counting of all consecutive codon pairs existing in the genome. It is used by the software to test the existence of non-association between two consecutive codons, through the Pearson chi-squared statistic, and to build the matrix of adjusted Pearson residual values which are associated to that statistical test [12]. The codon-pair context map corresponds to this matrix of adjusted Pearson residual values. For an easier visualization, each residual value, present in a cell of the contingency table, is converted into a two-color coded map. Green represents statistically significant positive values (associated to preferred codon pairs) and red represents statistically significant negative values (associated to rejected codon pairs) according to a predefined color scale. An illustration of a codon-pair context map obtained using the yeast's genome is presented in Fig. 3. The objective of codon-pair context maps is to detect patterns associated with preferred and rejected codon pairs. Applying biclustering algorithms, we

propose to contribute for the identification of forces that modulate codon-pair contexts and to identify additional patterns whose decoding by the ribosome might be highly problematic [3].

We initially observed that ISA's algorithmic strategy could be applied on codon-pair context maps of sequenced genomes. However, since the average is a central measure strongly influenced by errors and outliers in the data set, the current version of ISA publicly available at BicAT in http://www.tik.ee.ethz.ch/sop/bicat [9] may detect *undesirable* biclusters hiding relevant homogeneous groups. For instance, in the data matrix represented in Fig. 1(a), there is one very high value for each row g2−g5 and each column c4−c7 that yields abnormally high values for the averages for these rows and columns. Hence, it may be possible that ISA gives as outputs the submatrices represented in panel (b) of Fig. 1 hiding the true bicluster represented in panel (c), whose rows exhibit a similar pattern for all their columns. One way to overcome this problem is to modify ISA's criterion of biclustering using the statistical behavior of the median instead of the average.

The intervals $\mathcal{I}_x$ and $\mathcal{I}_y$ referred above claim for ISA the detection of biclusters whose rows (normalized and scored) have high values (we say, *Ygreater*) and whose columns (normalized and scored) have high absolute values (we say, *Xmodule*). When applied in microarray data matrices it is expected that ISA searches biclusters containing both up-regulated and down-regulated genes (i.e., highly negative and positive values in the same bicluster). We remark that in ISA while row scores are always positive, column scores can be positive and negative. If negative column scores are obtained in the first stage of one iteration of ISA, they will transform high negative values in high positive values in the second stage of the algorithm. Thus, ISA's strategy may not be adequate to identify biclusters containing only higher values or only lower values. ISA's bicluster structure can be entangled. This situation can easily be overcome considering unit scores or absolute scores, instead of scores with their signs, and incorporating other additional strategies on ISA's structure by replacing the predefined intervals $\mathcal{I}_x$ and $\mathcal{I}_y$ by more convenient ones. For instance, taking $\mathcal{I}_x = [\hat{\mu} - t_x\hat{\sigma}, +\infty[$, $\mathcal{I}_y = [\hat{\mu} - t_y\hat{\sigma}, +\infty[$ and absolute scores, the algorithm will search biclusters with low values by rows and columns (we say, *Xless*−*Yless*). For codon-pair context maps, we are interested in finding biclusters whose rows and columns contain only higher values (we say, *Xgreater*−*Ygreater*) and only lower values.

We investigated the effect of all those modifications in order to improve the performance of ISA. To do so, we constructed a new biclustering algorithm, herein called ISA-$Q_{\frac{1}{2}}$, based on ISA's structure using the statistical limiting behavior of the sample median and unit scores. Additionally, we replaced row and column scores by their absolute values in the original ISA and called this modified algorithm ISA-$|\overline{X}|$. Herein we provide a comparison and evaluation of the two new biclustering algorithms, ISA-$Q_{\frac{1}{2}}$ and ISA-$|\overline{X}|$, by opposition to the original ISA on different data sets and under three different combinations: (i) *Xmodule*−*Ygreater*, (ii) *Xgreater*−*Ygreater*, (iii) *Xless*−*Yless*. All these biclustering algorithms were implemented in Anaconda (available at http://www.bioinformatics.ua.pt/aplications/anaconda).

The remaining of the paper is organized as follows. In the next section, we describe ISA-$Q_{\frac{1}{2}}$. In Section 3, we proposed a methodology to understand and interpret the statistical relevance of biclusters in real number matrices which extends the definition of statistically significant biclusters in binary matrices given by Koyutürk [13]. Section 4 reports a comparative evaluation of the

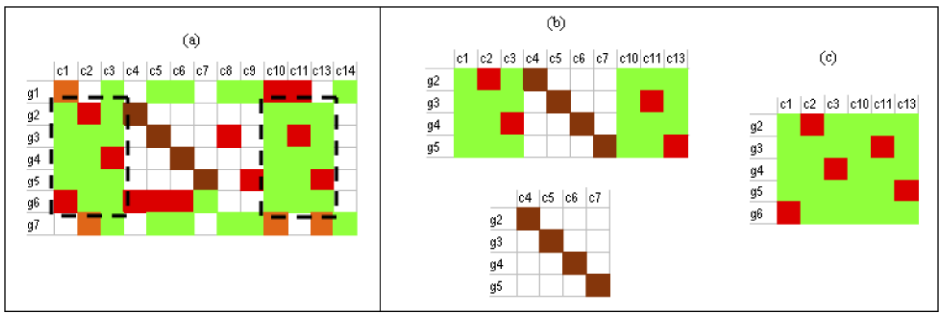

Fig. 1 Detection of biclusters from a matrix of real numbers with outliers (dark cells) and a relevant bicluster (dotted line) using ISA. (a) Data matrix with high values (red and green cells), low values (white cells) and a submatrix with extra high values or outliers in diagonal (dark cells). This submatrix is a potential bicluster to be detected by ISA, since each diagonal cell yields high averages by row and column. (b) Submatrices detected by ISA as biclusters. Both are not potentially significant in the sense that they are not 'unusually dense' (Section 3). (c) Statistically significant bicluster, where the rows exhibit similar behavior across the columns, and vice versa. The incorrect extra high values (dark cells) hinder ISA to detect the correct pattern in the data matrix independently of the threshold parameters used to filter the obtention of biclusters. Working with the median, instead of the average, the bicluster (c) can effectively be detected because the median is not affected by one extravagant observation

performance of the modifications of ISA referred above in two different perspectives: (i) with and (ii) without prior knowledge of biclusters implanted in data sets. In the last section, we summarize our main conclusions from our experimental studies.

2. ISA-MEDIAN

It is well known the following limit behavior of the sample median (see, for instance, Ref. 14, Theorem 10.5.1.)

THEOREM 1: Let $q_{\frac{1}{2}}$ be the median of a continuous random variable with density f. Let $(X_1, \ldots, X_s)$ be a sample from the distribution f and denote $Q_{\frac{1}{2}}$ the corresponding sample median. If f is positive and continuous at the point $x = q_{\frac{1}{2}}$, then $Q_{\frac{1}{2}}$ has the asymptotically Gaussian distribution given by

$$\mathcal{N}\Big(q_{\frac{1}{2}}, \big(\frac{1}{2\sqrt{s}f(q_{\frac{1}{2}})}\big)^2\Big), \qquad \text{as } s \to +\infty.$$

Taking into account the ISA algorithmic structure, this result allowed us to establish a similar biclustering criterion based on the behavior of the sample median, concretely, in terms of $Q_{\frac{1}{2}} - q_{\frac{1}{2}}$ compared with units of its standard deviation

$$\frac{1}{2\sqrt{s}f(q_{\frac{1}{2}})}\ . \tag{2}$$

The density probability distribution f depends on the probability distribution of the data that, in general, is unknown. We propose to consider $f(q_{\frac{1}{2}}) = \frac{1}{\sqrt{2\pi\sigma^2}}$ as if the data could effectively be fitted by a Gaussian distribution with mean μ and variance σ^2. Thus, Eq. (2) will be substituted by $\sigma\sqrt{\pi/2s}$. Getting t samples, we estimated $q_{\frac{1}{2}}$ by the average of the t medians. In order to obtain unbiased estimates, we estimated σ^2 by the variance of all observations in the data matrix $\mathbf{X}$ $\Big($i.e., $\widehat{\sigma} = \sqrt{\frac{\sum_{ij}(x_{ij}-\widehat{\mu})^2}{n\times m-1}}$ with $\widehat{\mu} = \frac{\sum_{ij} x_{ij}}{n\times m}\Big)$.

In the sequel, we describe the proposed steps of a single run of ISA-$Q_{\frac{1}{2}}$, where g is a function defined by $g(x, y) = x - y$, $g(x, y) = -(x - y)$ or $g(x, y) = |x - y|$, a choice which depends on the nature of the biclusters to be detected (high, low, or high absolute values, respectively), and $|A|$ denotes the number of elements in the set A. For ISA-$Q_{\frac{1}{2}}$, we considered a simplified structure of ISA, using non-normalized data matrices in their two stages and unit scores. This choice was motivated since it is simpler and, by this manner, each stage of the algorithm can be interpreted as the searching of rows and columns of the data matrix satisfying a pattern defined in terms of the behavior of theirs sample medians. The introduction of standardized data (having zero mean and unit variance), as used in ISA, would produce the same output as nonstandardized data, because unit scores are used.

Input:
$\mathbf{X}$: $n \times m$ matrix of the observations;
$C = \{C_j,\ j = 1, \ldots, m\}$—set of m columns;
$R = \{R_i,\ i = 1, \ldots, n\}$—set of n rows;
$R^{(0)}$—an initial set of $n_0 \leq n$ randomly selected rows;
t_y—threshold for the rows;
t_x—threshold for the columns.

First Stage

Step 1: Initialize $k = 0$.

Step 2: Obtain the submatrix of $\mathbf{X}$ for the selected rows $R^{(k)}$.

Step 3: Compute the medians by columns, S_{C_j}.

Step 4: Calculate the average of the medians by columns, $\overline{S}_C$.

Step 5: Obtain the subset $C^{(k)}$ of columns C_j satisfying a pattern defined by:

$$C^{(k)} = \left\{C_j \in C : g(S_{C_j}, \overline{S}_C) > t_x\sigma_C\right\}\ ,$$
$$\text{where}\quad \sigma_C = \widehat{\sigma}\sqrt{\frac{\pi}{2|R^{(k)}|}}.$$

Second Stage

Step 6: Obtain the submatrix of $\mathbf{X}$ for the selected columns $C^{(k)}$.

Step 7: Compute the medians by rows, S_{R_i}.

Step 8: Calculate the average of the medians by rows, $\overline{S}_R$.

Step 9: Obtain the subset $R^{(k+1)}$ of rows R_i satisfying a pattern defined by

$$R^{(k+1)} = \left\{R_i \in R : g(S_{R_i}, \overline{S}_R) > t_y\sigma_R\right\}\ ,$$
$$\text{where}\quad \sigma_R = \widehat{\sigma}\sqrt{\frac{\pi}{2|C^{(k)}|}}.$$

Step 10: If $R^{(k+1)} \neq R^{(k)}$ then make k equal to $k + 1$ and repeat Steps 2–9 else stop.

Output: Bicluster $= \left[x_{ij}\right]_{i\in R^{(k)},\ j\in C^{(k)}}$.

Each run of ISA-$Q_{\frac{1}{2}}$ yields at most one bicluster. In order to obtain more biclusters, eventually overlapped, the algorithm must be run several times.

Comparing with the original ISA, ISA-$Q_{\frac{1}{2}}$ presents various differences: (i) Steps 5 and 9, where the selection of rows and columns of biclusters is defined in terms of

medians and where the introduction of the function g allows the search of other kinds of biclusters, not necessarily taking the combination *Xmodule–Ygreater* (i.e., $g(x, y) = |x - y|$ in the first stage and $g(x, y) = x - y$ in the second stage) as proposed in ISA by its authors and implemented in BicAT; (ii) steps 2 and 6, where ISA-$Q_{\frac{1}{2}}$ considers always submatrices of $\mathbf{X}$ in opposition to ISA, which works, alternately, with submatrices of the standardized matrices by rows and columns of the matrix $\mathbf{X}$; (iii) steps 3 and 7, where there is no place for weights of rows and columns as there is in ISA. Note that, the function g can analogously be introduced in the original ISA. This was implemented in Anaconda.

3. STATISTICAL SIGNIFICANCE OF DISCOVERED BICLUSTERS

When a biclustering algorithm is applied on microarray data sets, the biological significance of each detected bicluster is usually analyzed checking its significant enrichment with respect to Gene Ontology (GO) annotations or other specific biological networks like metabolic and protein–protein interaction networks [11,15,16]. On a general real number matrix, how can we assign a bicluster as (statistically) significant? How do we proceed in order to determine the significance of a bicluster detected over a codon-pair context map?

In order to establish a criterion for the statistical significance of biclusters found by ISA, ISA-$Q_{\frac{1}{2}}$, and ISA-$|\overline{X}|$, which does not depend on any kind of biological relevance, we investigated how significant or 'unusually dense' these biclusters are, comparatively, with the initial matrix. The notion of an 'unusually dense' submatrix in a binary matrix was formalized by Koyutürk [13] and can be redefined in terms of one-side testing of hypothesis. Given an initial $n \times m$ binary matrix with k ones, a submatrix $\mathbf{B}$ is dense if it contains more ones than the initial matrix. Thus, the submatrix $\mathbf{B}$ is unusually dense, and so can be considered a *potentially significant* bicluster, if its observed number of ones leads to the rejection of the null hypothesis $H_{0,\mathbf{B}}$: $p_{\mathbf{B}} = p_0$ against the alternative hypothesis $H_{1,\mathbf{B}}$: $p_{\mathbf{B}} > p_0$, where $p_{\mathbf{B}}$ is the probability of finding ones in the submatrix $\mathbf{B}$ and $p_0 = k/nm$ is the proportion of ones in the initial matrix. For testing $H_{0,\mathbf{B}}$, we calculate the p-value in the following way:

$$p\text{-value}_{\mathbf{B}} = 1 - \phi\left(\frac{|1_{\mathbf{B}}|/|\mathbf{B}| - p_0}{\sqrt{\frac{p_0(1-p_0)}{|\mathbf{B}|}}}\right),$$

where $|1_{\mathbf{B}}|$ represents the number of ones in the submatrix $\mathbf{B}$, and $|\mathbf{B}|$ is the number of elements in $\mathbf{B}$ (=number of rows $\times$ number of columns). When $p\text{-value}_{\mathbf{B}} < \alpha$, the bicluster $\mathbf{B}$ will be classified as potentially significant at a level of significance α.

The biclustering methods ISA, ISA-$Q_{\frac{1}{2}}$, and ISA-$|\overline{X}|$ are described for real number matrices and identify no more than one bicluster for each application. Given a data matrix $\mathbf{X} = [x_{ij}]$, we investigated the statistical quality of biclusters sets obtained by a large number of runs of each biclustering algorithm, testing $H_{0,\mathbf{B}}$ for each identified bicluster $\mathbf{B}$. For that we take a discretization of $\mathbf{X}$ to a binary matrix $[b_{ij}]$, where

$$b_{ij} = \begin{cases} 1, & \text{se } g(x_{ij}, \widehat{\mu}) > \widehat{\sigma}\lambda \quad \text{and } \lambda \text{ is a} \\ 0, & \text{otherwise} \qquad\qquad\quad \text{threshold value} \end{cases}$$

where $g(x, y)$ depends on the strategy defined in the biclustering algorithm. Concretely, we took $g(x, y) = |x - y|$ for *Xmodule–Ygreater*, $g(x, y) = x - y$ for *Xgreater–Ygreater* and $g(x, y) = -(x - y)$ for *Xless–Yless* strategy.

Since this procedure may depend on the value of λ, it is recommended to execute the algorithm for different choices of λ. The nature of the data set can give a first natural suggestion for λ (Section 4). Steps 5 and 9 above considered may also suggest to select λ such that there is a certain percentage of observations x_{ij} in the data set satisfying the condition $g(x_{ij}, \widehat{\mu}) < \widehat{\sigma}\lambda$.

4. EXPERIMENTAL RESULTS

We have performed a comparative evaluation of the performance of ISA, ISA-$Q_{\frac{1}{2}}$, and ISA-$|\overline{X}|$ (i) with and (ii) without prior knowledge of implanted biclusters in data sets. For the first situation, we worked with the *in silico* data sets generated in Ref. 11, where, concerning the categories proposed by Madeira and Oliveira [10], two types of biclusters are implanted: (i) constant biclusters and (ii) additive biclusters and, concerning the structure, there are multiple biclusters not overlapping and with different overlap degrees. For the second situation, we analyzed real data matrices obtained with the organism *Saccharomyces cerevisiae* on two different approaches: (i) the codon-pair context map studied in Ref. 1, and (ii) the gene expression data set provided by Gasch *et al.* [17].

4.1. Having Prior Knowledge of Implanted Biclusters in Data Sets

The two artificial models used by Prelic *et al.* [11] provide synthetic data with biclusters defined by higher values in the data matrices. For both constant and additive

models, there are data matrices with noise and with non-overlapping and overlapping groups to investigate the sensitivity of each biclustering method to noise in the data and to overlapping in the biclusterings. In order to allow a fair comparison taking into account the original ISA, ISA-$|\overline{X}|$, and ISA-$Q_{\frac{1}{2}}$, we used the combination *Xmodule–Ygreater* and the parameter settings $t_x = t_y = 2$ as recommended by the authors of the original papers.

In order to assess the ability of each algorithm to recover known biclustering and reveal true grouping, we used the measure of match score defined by Lui and Wang [16].

DEFINITION 1: Let M_1 and M_2 be two sets of biclusters. The match score of M_1 with respect to M_2, herein denoted by $S_1(M_1, M_2)$, is equal to

$$\frac{1}{|M_1|}\sum_{(R_1,C_1)\in M_1}\max_{(R_2,C_2)\in M_2}\frac{|R_1\cap R_2|+|C_1\cap C_2|}{|R_1\cup R_2|+|C_1\cup C_2|},$$

where the pair (R, C) represents the submatrix whose rows and columns are given by the set R and C, respectively.

Let M_{opt} denote the set of implanted biclusters and M the set of the output of a biclustering algorithm. Thus, $S_1(M_{\text{opt}}, M)$ represents how well each of the true biclusters are detected by the algorithm under consideration. $S_1(M, M_{\text{opt}})$ quantifies how well each of the bicluster identified by the algorithm is represented in the set of true biclusters in both row and column dimensions. The measure S_1 is similar to the measure of match scores used in Ref. 11, but it has the advantage of reflecting, simultaneously, the match of the row and column dimensions between biclusters.

For the non-overlapping constant (additive) models, the used data sets correspond to ten 100×50 matrices of 0's and 1's (real numbers). Each matrix contains ten 10×5 implanted biclusters identified with higher values. These ten data matrices have various levels of noise. For each cell of the original data matrix, the noise was added and given by a random value drawn from a Gaussian distribution, where its standard deviation is the level of noise. For each noise level, ten data matrices were generated from each original data matrix and the averaged match score over these ten input matrices was calculated. For the overlapping constant (additive) models, the data sets correspond to nine $(100 + d) \times (100 + d)$, $d = 0, 1, 2, \ldots, 8$, matrices of 0's and 1's (real numbers), where d represents the overlap degree. Each matrix contains ten $(10 + d) \times (10 + d)$ implanted biclusters identified with higher values. For each overlap degree, the match score was calculated.

Figure 2 summarizes the performances of ISA, ISA-$|\overline{X}|$, and ISA-$Q_{\frac{1}{2}}$ with respect to the models considered. In the absence of noise, while the three algorithms are able to identify all implanted groups in the constant model ($S_1 = 100\%$), for the additive model the averaged match score decreases to 84% for ISA-$Q_{\frac{1}{2}}$ and holds in 100%

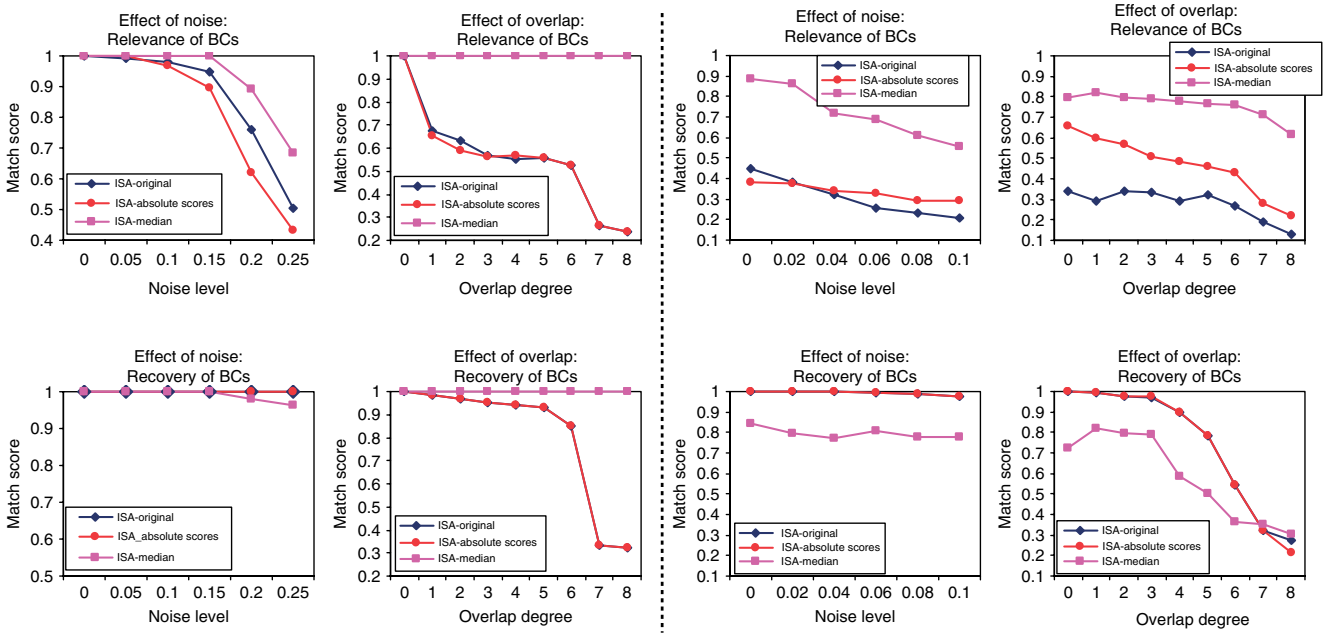

Fig. 2 Match score results for synthetic data where two types of biclusters are implanted in data matrices: constant biclusters (graphics on the left) and additive biclusters (graphics on the right), and under increasing noise level and increasing overlap degree. At the top, there are the values of $S_1(M, M_{\text{opt}})$ (relevance); at the bottom, there are the values of $S_1(M_{\text{opt}}, M)$ (recovery)

for the other two algorithms. In general, ISA-$|\overline{X}|$ showed better results than the original ISA except for revealing true biclusters implanted in the constant model. To reveal all true biclusters, ISA-$Q_{\frac{1}{2}}$ presented, comparatively, the best performance. ISA-$Q_{\frac{1}{2}}$ only presented worse performance in recovering all implanted groups for additive models (averaged match scores around 80% in the presence of noise and smaller than 80% when the true biclusters are overlapped). In fact, ISA-$Q_{\frac{1}{2}}$ exhibits a general tendency to find fewer biclusters and, therefore, it will be less probable to identify all true biclusters (recovery, Fig. 2 at the bottom); nevertheless, these found biclusters are more probable to be true biclusters in contrast to ISA and ISA-$|\overline{X}|$ which exhibit tendency to identify several non-true biclusters (revelance, Fig. 2 at the top). Since the robustness of the median, it is more probable that true biclusters can be picked up by ISA-$Q_{\frac{1}{2}}$, whether in the presence of noise or when true biclusters are overlapped. In opposition, outcomes from the biclustering algorithms based on the average can be quite influenced by the presence of noise and overlapping.

4.2. Having No Prior Knowledge of Biclusters Implanted in Data Sets

Next, we focused our attention on real data sets where the existence of biclusterings is unknown. We addressed our study over two different data sets from the organism *S. cerevisiae*. The first one is the codon-pair context map of the *S. cerevisiae*. It is the 64×64 real data matrix obtained by Anaconda software [2] after reading and interpreting the total coding sequences of all the chromosomes of *S. cerevisiae* downloaded from the National Center for Biotechnology Information ftp site (ftp://ftp.ncbi.nih.gov/genomes/). The second data set is a microarray data matrix which contains 2993 genes of the *S. cerevisiae* over 173 different stress conditions. This gene expression data set was provided by Gasch *et al.* [17].

We analyzed different combinations: (i) *Xmodule–Ygreater*; (ii) *Xgreater–Ygreater*; and (iii) *Xless–Yless* for each one of the three algorithms ISA, ISA-$|\overline{X}|$, and ISA-$Q_{\frac{1}{2}}$. In order to assess the ability of each algorithm to recover biclusters detected by others, we calculated the match scores $S_1(M_1, M_2)$, for $M_1 \neq M_2$ and $M_1, M_2 = M_{\overline{X}}, M_{|\overline{X}|}, M_{Q_{\frac{1}{2}}}$, where $M_{\overline{X}}$, $M_{|\overline{X}|}$, and $M_{Q_{\frac{1}{2}}}$ denotes the set of biclusters detected by ISA, ISA-$|\overline{X}|$, and ISA-$Q_{\frac{1}{2}}$, respectively. Note that $S_1(M_A, M_B)$ quantifies how well each bicluster identified by algorithm A, is also detected by algorithm B. However, if there are biclusters detected by one algorithm contained in bigger biclusters detected by the other algorithm, the measure S_1 does not show it. To analyze this situation, we also calculated a second measure match score given as follows.

DEFINITION 2: Following notation of Definition 1, $S_2(M_1, M_2)$ is equal to

$$\frac{1}{|M_1|} \sum_{(R_1,C_1)\in M_1} \max_{(R_2,C_2)\in M_2} \frac{|R_1 \cap R_2| + |C_1 \cap C_2|}{|R_1| + |C_1|}.$$

$S_2(M_A, M_B)$ quantifies how well each bicluster identified by algorithm A is contained into some bicluster detected by algorithm B.

4.2.1. Yeast codon context data set

We analyzed the capability of ISA, ISA-$|\overline{X}|$, and ISA-$Q_{\frac{1}{2}}$ on the detection of general patterns of codon-pair contexts in sequenced genomes when applied on the codon-pair context map of *S. cerevisiae*. This data matrix is illustrated in Fig. 3 (on the left) and was obtained using Anaconda software. The main goal was the identification of patterns associated to preferred and rejected codon pairs.

In a first experimental evaluation, we observed that the three algorithms allowed to identify potentially significant biclusters, in the sense given in Section 3 with $\alpha = 0.05$ and $\lambda = 3$, and can identify patterns not detected using classical hierarchical algorithms. In general, the significant biclusters detected by ISA-$Q_{\frac{1}{2}}$ were bigger for many combinations of t_x, t_y. In Fig. 3 (on the right), two significant biclusters detected by ISA-$Q_{\frac{1}{2}}$ are presented. One pattern (the topmost one) is inline with previous results by Moura *et al.* [1]. The other was a new result from this algorithm.

On the other hand, it seemed natural that the value λ could produce differences in the estimation of the statistical significance of discovered biclusters. To investigate this fact, we analyzed the impact of λ for the algorithms ISA, ISA-$|\overline{X}|$, and ISA-$Q_{\frac{1}{2}}$, when $t_x = t_y = 2$ and for the combinations *Xmodule–Ygreater* and *Xgreater–Ygreater* (for ISA *Xless–Yless*, potentially significant biclusters were not detected, cf. Fig. 5). For such ten replicates of 500 runs of each method were constructed and the percentage of potentially significant biclusters with $\lambda = 1, 2, \ldots, 9$, at a level of significance $\alpha = 0.05$, was calculated for each replication. Comparative boxplots (Fig. 4) for the observed percentages showed that for *Xmodule–Ygreater* the percentage of potentially significant biclusters identified by ISA-$Q_{\frac{1}{2}}$ is higher independently of the value of λ. For *Xgreater–Ygreater*, $\lambda = 5$ provided lower percentages of a bicluster detected by ISA-$Q_{\frac{1}{2}}$ being potentially significant comparatively with the other two algorithms. We decided to fix the same $\lambda = 3$ for the discretization of the codon-pair

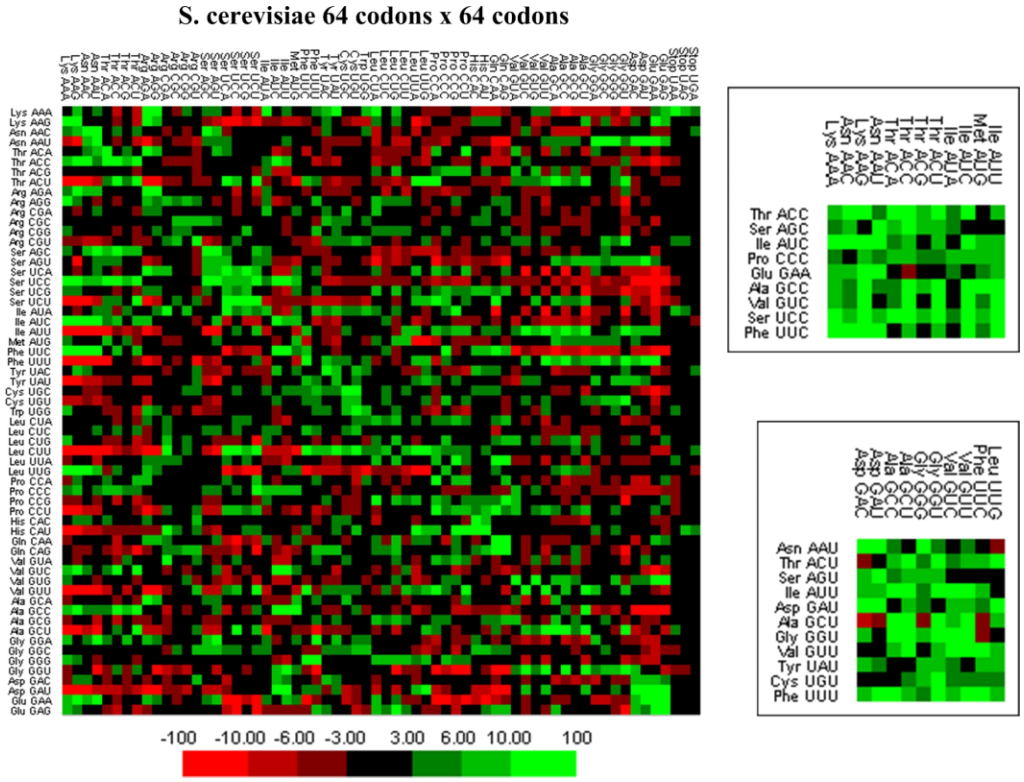

Fig. 3 Codon-pair context map of *S. cerevisiae* (left) and two potentially significant biclusters (right), $\alpha = 0.05$ and threshold value $\lambda = 3$, obtained by ISA-$Q_{\frac{1}{2}}$, with $t_x = t_y = 2$, where the patterns NNC-ANN and NNU-GNN stood out as highly preferred in the genome of *S. cerevisiae*. Here N represents any nucleotide A, C, G, or U

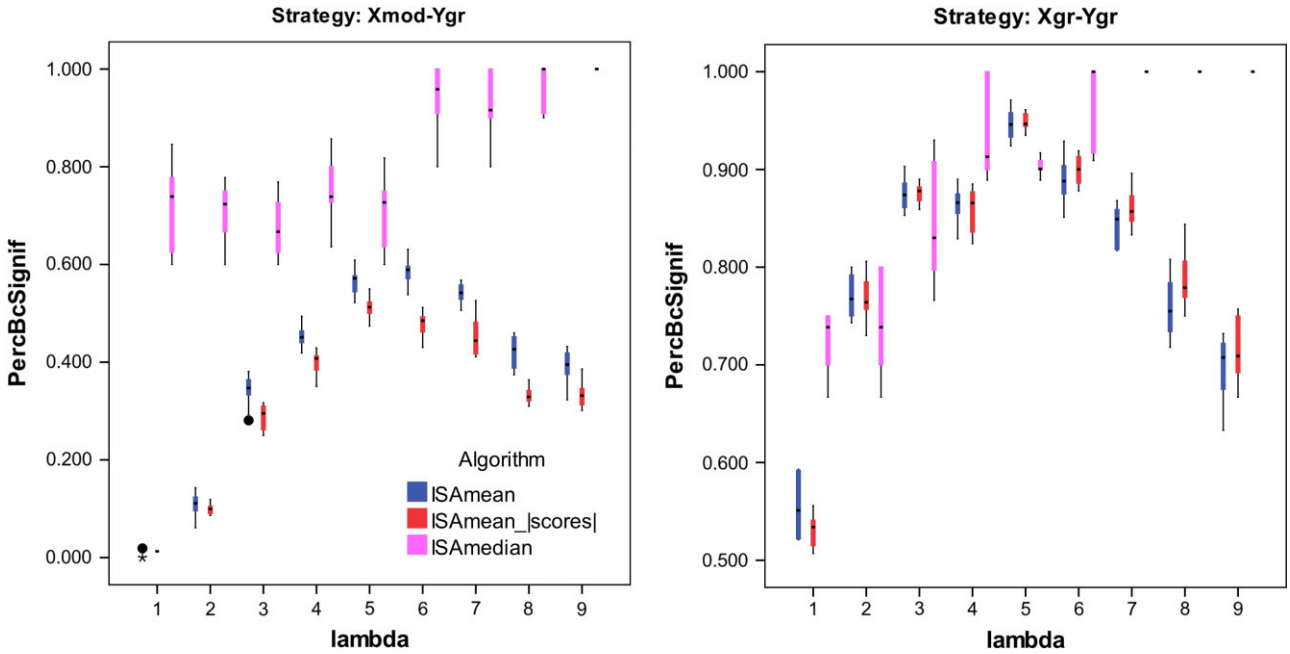

Fig. 4 Boxplots of the empirical distribution of the percentage of a bicluster detected by ISA, ISA-$|\overline{X}|$, and ISA-$Q_{\frac{1}{2}}$ (*Xmodule–Ygreater*—left—and *Xgreater–Ygreater*—right), with $t_x = t_y = 2$, being potentially significant at a level of significance $\alpha = 0.05$ when the parameter of discretization is $\lambda = 1, 2, \ldots, 9$. For ISA-$Q_{\frac{1}{2}}$, the percentage decreases from $\lambda = 19$

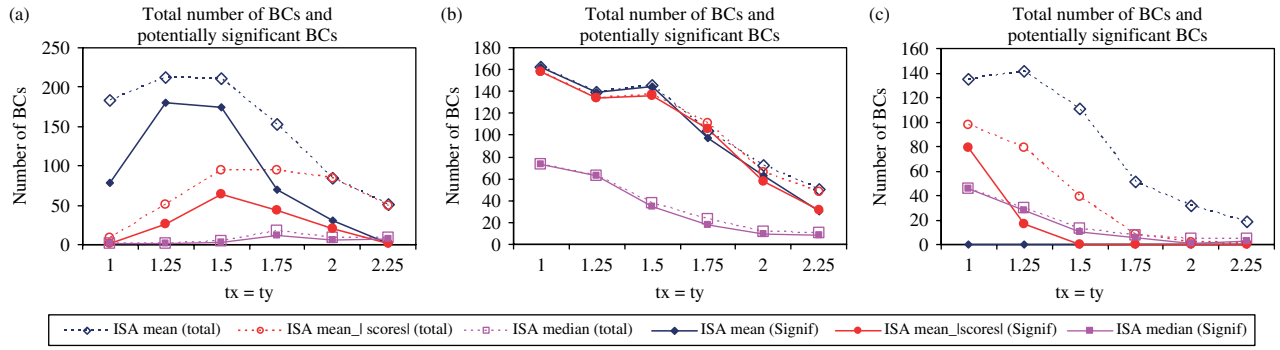

Fig. 5 Total number of distinct biclusters (dashed line) and number of potentially significant biclusters (solid line) formed by running 500 times ISA (blue lines), ISA-$|\overline{X}|$ (red lines), and ISA-$Q_{\frac{1}{2}}$ (pink lines) on the codon-pair context map of *S. cerevisiae* for the combinations (a) *Xmodule–Ygreater*, (b) *Xgreater–Ygreater*, and (c) *Xless–Yless*. Their dependence on the threshold parameters $t_x = t_y$. The distance between two lines of the same color indicates how many biclusters identified by one algorithm are not potentially significant. For the case (c) all biclusters detected by ISA are not potentially significant for any choice of $t_x = t_y$

context map of *S. cerevisiae* in the evaluation of the three biclustering algorithms on that data set. This choice was intuitively suggested by the conversion used to color the codon-pair context map [1].

We provided a quantitative analysis of the performance of the three biclustering methods based on (i) the number of biclusters and potentially significant biclusters found by each algorithm (Fig. 5), and (ii) the match scores $S_i(M_1, M_2)$, $i = 1, 2$ for $M_1 \neq M_2$ and $M_1, M_2 = M_{\overline{X}}, M_{|\overline{X}|}, M_{Q_{\frac{1}{2}}}$ (Fig 6). For these computations, each algorithm was run 500 times on the codon-pair context map of *S. cerevisiae* for several combinations of the threshold parameters t_x and t_y and found biclusters were classified as potentially significant according to Section 3, with $\lambda = 3$ and $\alpha = 0.05$. Since one goal is the identification of patterns of higher values, all the algorithms should be applied taking $g(x - y) = x - y$ in their two stages. Nevertheless, we analyzed three different situations for the function g: (i) $g(x, y) = |x - y|$ in first stage and $g(x, y) = x - y$ in second stage (i.e., *Xmodule–Ygreater*); (ii) $g(x, y) = x - y$ in

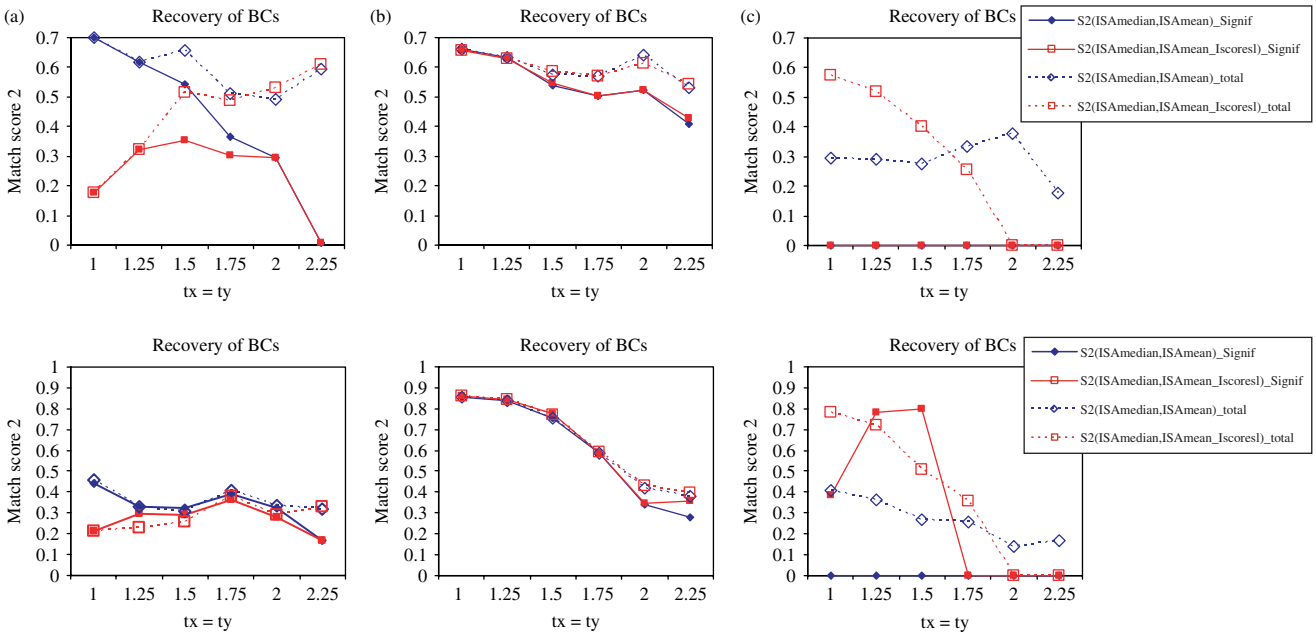

Fig. 6 Measures of match scores S_2, of $M_{Q_{\frac{1}{2}}}$ with respect to $M_{\overline{X}}$ and $M_{|\overline{X}|}$ (top) and of $M_{\overline{X}}$ and $M_{|\overline{X}|}$ with respect to $M_{Q_{\frac{1}{2}}}$ (bottom), depending on the threshold parameters $t_x = t_y$. The sets of all the biclusters found (*total*) and the sets of all the potentially significant biclusters (*Signif*) detected by 500 runs of each algorithm are considered

both stages (i.e., *Xgreater–Ygreater*); and (iii) $g(x, y) = -x + y$ in both stages (i.e., *Xless–Yless*). The outcomes of the biclustering algorithms for each combination are available at http://bioinformatics.ua.pt.

In general, ISA yields a greater number of outputs. Nevertheless, obtaining an increasing number of biclusters does not imply that they are more significant biclusters. Figure 5 depicts how many found biclusters are potentially significant. While for any combination and for any parameters $t_x = t_y$, the biclusters detected by ISA-$Q_{\frac{1}{2}}$ are, in general, potentially significant, it does not hold for the other algorithms, except for the combination *Xgreater–Ygreater*. For ISA-$Q_{\frac{1}{2}}$, the parameters $t_x = t_y$ presented less dependence on the total number of detected biclusters. For the case *Xless–Yless*, ISA showed to be inadequate since none of the found biclusters are potentially significant. This unexpected result is due to the fact that the original ISA allows for the use of negative row and column scores. Negative scores change negative high values eventually existing in rows and columns into positive high values and hence allow for the selection of these rows and columns and its detection as a bicluster. By taking absolute scores this effect is eliminated. In that case, ISA should be substituted by ISA-$|\overline{X}|$ or ISA-$Q_{\frac{1}{2}}$.

We also evaluated the capability of one biclustering algorithm to recover the biclusters detected by others through the measures of match scores S_1 and S_2. To illustrate this approach, Fig. 6 schematizes the values obtained for S_2. Results obtained for S_1 leads to analogous conclusions and were therefore omitted. The results of S_2 reflect a higher capability of ISA-$Q_{\frac{1}{2}}$ to reveal large biclusters containing biclusters detected by ISA and ISA-$|\overline{X}|$. Indeed, while the three graphics at the top in Fig. 6 depict the ability of ISA-$Q_{\frac{1}{2}}$ to reveal biclusters contained in biclusters also detected by the other two algorithms, the three graphics at the bottom exhibit the capability of ISA and ISA-$|\overline{X}|$ to reveal biclusters that were included in bigger biclusters identified by ISA-$Q_{\frac{1}{2}}$. Comparatively, and in many cases, these last graphics present higher match scores, showing that ISA-$Q_{\frac{1}{2}}$ has a tendency to yield fewer and bigger biclusters. This characteristic was also verified when constant and additive biclusters are implanted in the data matrix (Section 4.1). Furthermore, we remark that those higher S_2 scores are not random artifact due to larger bicluster size. We analyzed the combination *Xgreater–Ygreater* for $t_x = t_y = 1$, where the highest match scores were obtained: $S_2(M_{\overline{X}}, M_{Q_{\frac{1}{2}}}) = 0.853$, $S_2(M_{Q_{\frac{1}{2}}}, M_{\overline{X}}) = 0.659$, $S_2(M_{|\overline{X}|}, M_{Q_{\frac{1}{2}}}) = 0.859$, and $S_2(M_{Q_{\frac{1}{2}}}, M_{|\overline{X}|}) = 0.658$ (cf. Fig. 6). In order to show the true significance of those S_2 scores, biclusters obtained by ISA-$Q_{\frac{1}{2}}$ were substituted by random biclusters (i.e., biclusters of same size of the formers but with the rows and columns randomly generated). Twenty replications were executed defining 20 sets M of the random biclusters. Calculations of the match scores mentioned above using the generated sets M instead of $M_{Q_{\frac{1}{2}}}$ leaded to the averaged match scores $S_2(M_{\overline{X}}, M_{Q_{\frac{1}{2}}})$, $S_2(M_{|\overline{X}|}, M_{Q_{\frac{1}{2}}}) \approx 0.51$ and $S_2(M_{Q_{\frac{1}{2}}}, M_{\overline{X}})$, $S_2(M_{Q_{\frac{1}{2}}}, M_{|\overline{X}|}) \approx 0.34$, with standard deviations ≈ 0.01. Using both the t-test and Wilcoxon signed-rank test, we obtained a p-value $= 0.000$ of obtaining similar match scores S_2 when a set of random biclusters or the set $M_{Q_{\frac{1}{2}}}$ are considered.

Moreover, from Fig. 6, there is a high difference between $S_2(M_{Q_{\frac{1}{2}}}, M_{\overline{X}})$ and $S_2(M_{Q_{\frac{1}{2}}}, M_{|\overline{X}|})$, for some threshold parameters $t_x = t_y$. This indicates that biclusters found by ISA-$Q_{\frac{1}{2}}$ are revealed by ISA and ISA-$|\overline{X}|$ in distinct ways. We emphasize the combinations *Xgreater–Ygreater* and *Xless–Yless* (panels (b) and (c) of Fig. 6). For the case (b) and for $t_x = t_y = 1, 1.25, 1.5$, where there was the highest larger number of biclusters found by the three algorithms (cf. Fig. 5), we observed that $S_2(M_{\overline{X}}, M_{Q_{\frac{1}{2}}})$ and $S_2(M_{|\overline{X}|}, M_{Q_{\frac{1}{2}}})$ show that >70% of all significant biclusters formed by ISA were recovered by bigger significant biclusters identified by ISA-$Q_{\frac{1}{2}}$ which contained the first. In contrast, $S_2(M_{Q_{\frac{1}{2}}}, M_{\overline{X}})$ is less than 70%, indicating less capability for ISA. Also, for the case (c), ISA-$Q_{\frac{1}{2}}$ presented a better performance for $t_x = t_y = 1, 1.25$ for which there was the highest number of biclusters found by the three algorithms. Effectively, the introduction of the function g and the sample median in ISA's structure allowed to unveil more adequate biclusters on the codon-pair context map of *S. cerevisiae*. Therefore, to identify patterns of preferred and rejected codon pairs on the codon-pair context map of any species, we recommend to consider ISA-$Q_{\frac{1}{2}}$ with $g(x, y) = x - y$ and $g(x, y) = -(x - y)$, respectively, in the two stages of the algorithm.

4.2.2. *Yeast expression data set*

For the yeast expression data given in Ref. 17, ISA-$Q_{\frac{1}{2}}$ revealed to be inefficient. This data set hindered this algorithm from achieving the stopping criterion (Step 10) or leaded to the find of few and big biclusters, particularly for the combinations *Xgreater–Ygreater* and *Xless–Yless*. In opposition, using the algorithms based on sample averages, ISA and ISA-$|\overline{X}|$, many transcription modules were detected. How meaningful are these biclusters? We are particularly interested in analyzing the influence of the use of negative scores in ISA's strategy. Thus, a comparative study was carried out considering the original ISA (i.e., ISA with combination *Xmodule–Ygreater* and row and column scores with their signs) and ISA-$|\overline{X}|$ for the

same combination, as reference algorithms. Therefore, for each combination *Xmodule–Ygreater*, *Xgreater–Ygreater*, and *Xless–Yless*, we computed $S_i(bs, M)$ and $S_i(M, bs)$, $i = 1, 2$, where M and bs represents the set of all biclusters detected by one algorithmic strategy and one reference algorithm, respectively. To obtain M, for each algorithm ISA and ISA-$|\overline{X}|$ and for each combination, 500 runs of each algorithm was executed with $t_x = t_y = 2$. For the classification of each found bicluster as potentially significant, we took $\lambda = 1$ for the discretization and a level of significance $\alpha = 0.001$ (cf. Section 3). The choice of this value of the parameter λ was empirical and consequence of some properties observed for the distribution of this data (mean = 0.19, median = 0, quasy-symmetric and the most central part (86%) of gene expression levels are observed between -1 and 1.5).

Firstly, we considered as reference the set of all biclusters detected by the original ISA (i.e., ISA with combination *Xmodule–Ygreater* and row and column scores with their signs). In a second analysis, the reference was the set of all biclusters detected by ISA-$|\overline{X}|$ with combination *Xmodule–Ygreater*. The obtained match scores S_2 are shown in Fig. 7. Similar behavior was obtained for S_1 (data not shown). When the reference is the original ISA (blue lines in Fig. 7), the values of the measures of match scores when M is the set of biclusters in the combination *Xless–Yless* (for instance, when M resulted from ISA-$|\overline{X}|$ we obtained: $S_1(bs, M) = 0.391$, $S_1(M, bs) = 0.626$, $S_2(bs, M) = 0.459$, $S_2(M, bs) = 0.726$) demonstrate ability of the original ISA to recover and reveal biclusters with lower values but not all biclusters in that condition that were detected by ISA-$|\overline{X}|$. For the combination *Xgreater–Ygreater*, the calculations of the match scores (all $\approx$0.50) lead to similar conclusions for biclusters with higher values. When the reference biclustering algorithm is ISA-$|\overline{X}|$ (red lines in Fig. 7) the measures of match scores $S_1(bs, M)$ and $S_2(bs, M)$ exhibited high values ($\geq$0.89) when M corresponded to the combination *Xgreater–Ygreater* and low values ($\leq$0.16) for the *Xless–Yless* situation. These results indicate a high and low ability of the combination *Xgreater–Ygreater* and *Xless–Yless*, respectively, for recovering the bicluster detected by that reference. This conclusion is consistent with the strategy algorithm associated to these combinations.

Selecting the combination *Xmodule–Ygreater*, we assessed the ability of ISA and ISA-$|\overline{X}|$ to find biologically relevant biclusters on the microarray data set. For such, we explored how the biclusters are significantly enriched in GO annotations. For each detected bicluster, we used FuncAssociate software [18] to obtain the adjusted p-value associated with each GO term existing on the bicluster's gene list and compute the proportion of biclusters significant enrichment in GO annotations. Also, the number of attributes significantly over-represented, at levels of significance $\alpha = 0.0001, 0.001, 0.005, 0.01, 0.05$, was retained.

Both algorithms, original ISA and ISA-$|\overline{X}|$, provided a high percentage of biclusters containing genes enriched in GO annotations at all levels of significance considered (Fig. 8), having the original ISA generated 122 biclusters while ISA-$|\overline{X}|$ generated 69 biclusters. Using the chi-squared Pearson statistics test, while there is no statistically

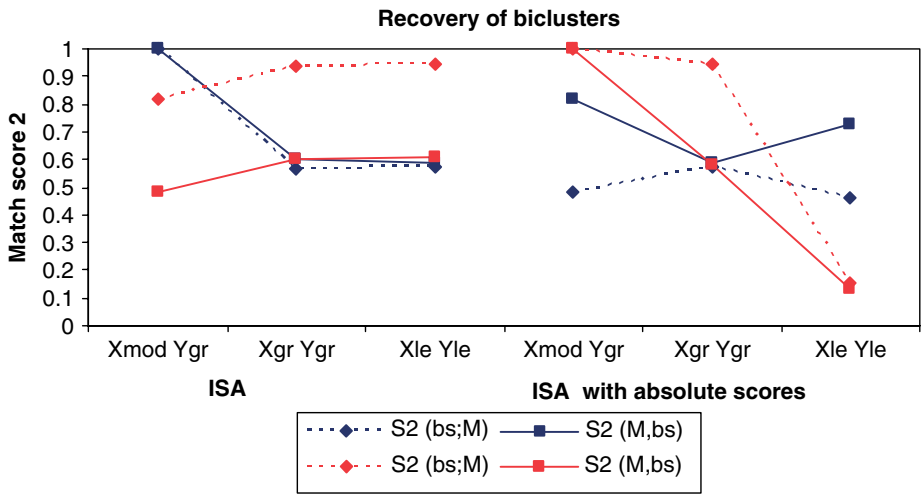

Fig. 7 Performance of ISA applied to gene expression data matrix of the organism *S. cerevisiae*. Values of the match scores S_2 of M with respect to a reference biclustering algorithm bs, $S_2(M, bs)$—solid lines—and vice versa, $S_2(bs, M)$—dotted lines—are represented. While for blue lines the reference is the original ISA (i.e., ISA with scores without module and combination *Xmodule–Ygreater*), for red lines the reference is ISA-$|\overline{X}|$ and the same combination *Xmodule–Ygreater*. M represents the output of each biclustering method indicated in the axis x

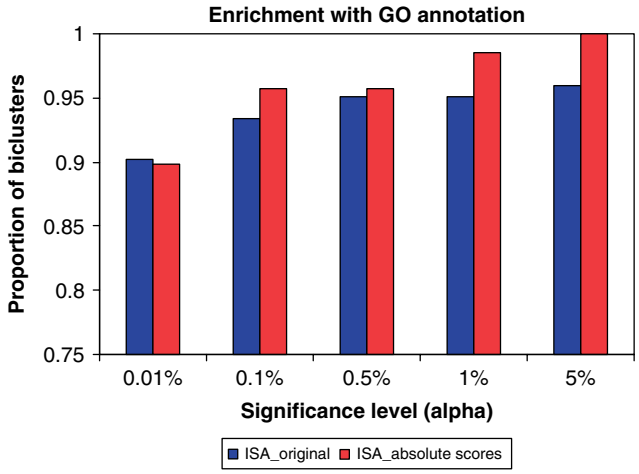

Fig. 8 Proportion of biclusters significantly enriched in GO annotation on *S. cerevisiae*'s gene expression data set, at different levels of significance α, for the original ISA (blue bars) and ISA-$|\overline{X}|$ (red bars)

significant association between a bicluster being potentially significant and being enriched in GO annotations, at a level of significance $\alpha = 0.0001$ (p-value ≥ 0.490), for $\alpha = 0.05$ this conclusion is not so clear (p-value ≥ 0.052). Figure 9 shows how the quantity of significantly over-represented GO terms were distributed in potentially significant biclusters detected by both algorithms. The curve delineated by red points shows more abrupt increasing than the one by blue points, namely around $a = 0.1, 0.2$ for $\alpha = 0.0001$ and $a = 0.4, 0.5$ for $\alpha = 0.05$. This means that ISA-$|\overline{X}|$ detected a lower percentage of potentially significant biclusters with a low number of over-represented attributes.

5. CONCLUSION

We analyzed in detail the biclustering method ISA, pointed its main fragilities and proposed procedures in order to improve its performance. Assuming unit scores, we could say that ISA will search biclusters whose rows and columns belong to critical regions of statistical tests for the mean defined in terms of z-score statistics. Consequently, extensions of ISA's structure for other types of statistics can be developed. Herein, the median instead of the average is proposed. Modifications into ISA's structure were then explained leading to the description of the algorithms ISA-$|\overline{X}|$ and ISA-$Q_{\frac{1}{2}}$ with the possibility of the identification of biclusters with high values (combinations $Xmodule-Ygreater$ and $Xgreater-Ygreater$) and low values ($Xless-Yless$). A comparative empirical study of the performance of the three biclustering algorithms for these three combinations is herein reported in a detailed and systematic way using both synthetic and real data sets. Our experiments show that ISA-$Q_{\frac{1}{2}}$ outperforms ISA in most cases, namely (i) it is more resilient to the outcome of biclusters without significance; (ii) in general, it recovers, with high percentage, all implanted biclusters in data sets; (iii) its capability for revealing all true biclusters appears to be less sensitive to noise in the data and to overlapping degree in groups; (iv) the input parameters have less impact on its performance; (v) the resulting biclusters have a greater tendency to be potentially significant than the biclusters discovered by ISA. In general, ISA-$|\overline{X}|$ presented better performance than ISA. In many cases, ISA-$Q_{\frac{1}{2}}$ outperformed ISA-$|\overline{X}|$ showing a higher tendency to find fewer and bigger biclusters than the other two methods. The biclusters detected by ISA-$Q_{\frac{1}{2}}$ are more probable to be true biclusters in contrast to ISA and

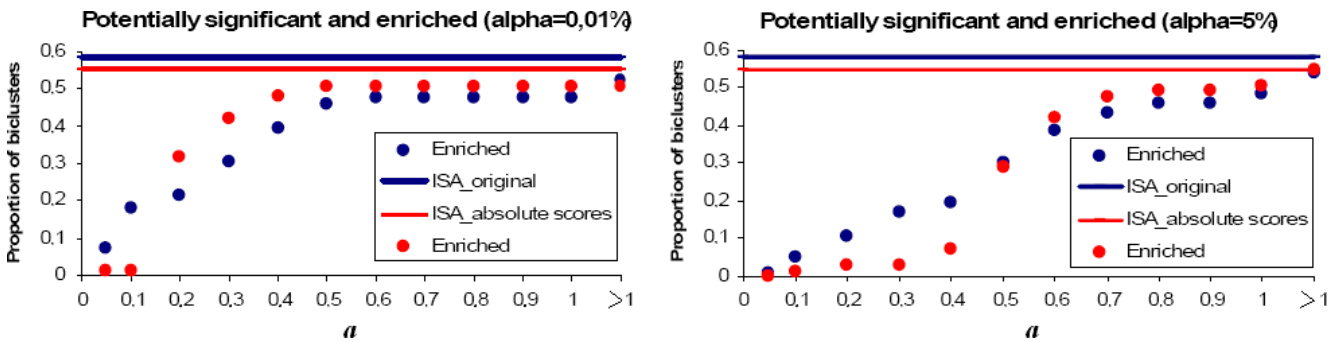

Fig. 9 Potentially significant and enriched in GO annotations biclusters detected on *S. cerevisiae*'s gene expression data set. Proportion of potentially significant biclusters, at a level of significance of 5%, for each biclustering method (horizontal lines), and the proportion of biclusters enriched in GO annotations, at different levels of significance $\alpha = 0.0001, 0.05$, containing $a \times$ (number of genes belonging to each bicluster) $* 100\%$ of over-represented attributes (solid dots). While results for the original ISA are represented in blue color, for ISA-$|\overline{X}|$ they are in red. The differences between the latter points ($a > 1$) and the horizontal lines mean there are potentially significant biclusters detected by both algorithms which have no GO terms over-represented in genes belonging to them

ISA-$|\overline{X}|$ which exhibit tendency to identify several non-true biclusters.

ACKNOWLEDGMENTS

We are indebted to Dorabella Santos and Teresa for correcting the English. We also thank the anonymous referees for their constructive comments which improved this work. This research was supported by *Fundação para a Ciência e a Tecnologia* (Portugal) through PTDC/MAT/72974/2006 and Center of Research and Development in Mathematics and Applications of University of Aveiro.

REFERENCES

[1] G. Moura, M. Pinheiro, R. Silva, I. Miranda, V. Afreixo, G. Dias, A. Freitas, J. L. Oliveira, and M. Santos, Comparative context analysis of codon pairs on an ORFeome scale, Genome Biol 6 (2005), R28.

[2] M. Pinheiro, V. Afreixo, G. Moura, A. Freitas, M. A. Santos, and J. L. Oliveira, Statistical, computational and visualization methodologies to unveil gene primary structure features, Meth Inform Med 45 (2006), 163–168.

[3] G. Moura, M. Pinheiro, J. Arrais, A. C. Gomes, L. Carreto, A. Freitas, J. L. Oliveira, and M. Santos, Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure, PLoS ONE 2(9) (2007), e847, doi:10.1371/journal.pone.0000847.

[4] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, and Y. Ziv, Revealing modular organization in the yeast transcriptional network, Nat Genet 31 (2002), 370–377.

[5] M. B. Eisen, P. T. Spellman, P. Brown, and D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc Natl Acad Sci USA 95 (1998), 14863–14868.

[6] S. Datta and S. Datta, Comparisons and validation of statistical clustering techniques for microarray gene expression data, Bioinformatics 19 (2003), 459–466.

[7] Y. Cheng and G. Church, Biclustering of expression data, Proc Int Conf Intell Syst Mol Biol 8 (2000), 93–103.

[8] J. Ihmels, S. Bergmann, and N. Barkai, Defining transcription modules using large-scale gene expression data, Bioinformatics 20 (2004), 1993–2003.

[9] S. Barkow, S. Bleuler, A. Prelic, P. Zimmermann, and E. Zitzler, BicAT: a biclustering analysis toolbox, Bioinformatics 22 (2006), 1282–1283.

[10] S. Madeira and A. Oliveira, Biclustering algorithms for biological data analysis: a survey, IEEE/ACM Trans Comput Biol Bioinform 1 (2004), 24–45.

[11] A. Prelic, 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 (2006), 1122–1129.

[12] A. Agresti, Categorical Data Analysis (2nd ed.), Wiley, NY, 2002.

[13] M. Koyutürk, W. Szpankowski, and A. Grama, Biclustering gene-features matrices for statistically significant dense patterns, In Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference, 2004, 480–484.

[14] A. Fisz, Probability Theory and Mathematical Statistics, (3rd ed.), Krieger, Florida, 1963.

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

[16] X. Lui and L. Wang, Computing the maximum similarity biclusters of gene expression data, Bioinformatics 23 (2007), 50–56.

[17] A. Gasch, P. Spellman, C. Kao, O. Carmel-Harel, M. Eisen, G. Storz, D. Botstein, and P. Brown, Genomic expression programs in the response of yeast cells to environmental changes, Mol Biol Cell 11 (2000), 4241–4257.

[18] G. Berriz, O. King, B. Bryant, C. Sander, and F. Roth, Characterizing gene sets with FuncAssociate, Bioinformatics 19 (2003), 2502–2504.