# Proximity Measures for Clustering Gene Expression Microarray Data: A Validation Methodology and a Comparative Analysis

Pablo A. Jaskowiak, Ricardo J.G.B. Campello, and Ivan G. Costa

**Abstract**—Cluster analysis is usually the first step adopted to unveil information from gene expression microarray data. Besides selecting a clustering algorithm, choosing an appropriate proximity measure (similarity or distance) is of great importance to achieve satisfactory clustering results. Nevertheless, up to date, there are no comprehensive guidelines concerning how to choose proximity measures for clustering microarray data. Pearson is the most used proximity measure, whereas characteristics of other ones remain unexplored. In this paper, we investigate the choice of proximity measures for the clustering of microarray data by evaluating the performance of 16 proximity measures in 52 data sets from time course and cancer experiments. Our results support that measures rarely employed in the gene expression literature can provide better results than commonly employed ones, such as Pearson, Spearman, and euclidean distance. Given that different measures stood out for time course and cancer data evaluations, their choice should be specific to each scenario. To evaluate measures on time-course data, we preprocessed and compiled 17 data sets from the microarray literature in a benchmark along with a new methodology, called Intrinsic Biological Separation Ability (IBSA). Both can be employed in future research to assess the effectiveness of new measures for gene time-course data.

**Index Terms**—Proximity measure, distance, similarity, correlation coefficient, clustering, gene expression, cancer, time course

## 1 Introduction

Microarray technology enables expression level measurement for thousands of genes in a parallel fashion. The genomic picture obtained with microarrays helps researchers to gather knowledge and insights on diverse biological phenomena. However, the same huge amount of data that bring excitement introduces big challenges regarding its interpretation and analysis. To tackle these challenges, diverse computational methods have been developed and employed. Among these, clustering plays an important role, as it is one of the first steps adopted to unveil information from gene expression data [1], [2].

Clustering has two major applications, depending on the type of microarray experiments under analysis. The first is found when expression of genes is monitored across time for a biological process of interest. In the so-called time-course experiments, clustering may help, for instance, to identify genes that share the same regulatory mechanisms or functions [3], [4], [5], [6]. The second application involves the analysis of biological samples, usually from different types of cancer. The main interest in this scenario lies in detecting previously unknown clusters of patient samples, i.e., clusters of previously unknown types of cancer [7], [8], [9], [10].

A great variety of clustering algorithms has been developed specifically for the clustering of genes (e.g., [4], [11], [12], [13]) and cancer samples (e.g., [2], [14], [15], [16]). Simultaneously, diverse classical clustering algorithms, such as the well-known k-means [17], [18] and hierarchical methods [19] have been frequently used in both scenarios. Due to the importance of choosing an appropriate clustering algorithm, numerous theoretical reviews and empirical comparisons regarding the clustering of genes [20], [21], [1], [5], [22], [6] and cancer samples [1], [6], [23], [24], [25] have been presented.

Albeit important, the choice of the clustering algorithm itself is not the only determining factor for clustering results and their quality. As a matter of fact, the choice of an appropriate proximity measure (similarity or distance[1]) employed between object pairs is often regarded as a central issue in clustering analysis [27], [5], [26]. Studies have compared proximity measures for different application domains, such as text clustering [28], ontology annotations [29], and clustering-based feature selection [30], providing guidelines for each one.

Despite the wide variety of proximity measures described in the clustering literature, a particular measure is usually preferred given the characteristics of the problem in hand [5], [27], [26]. For gene expression data, it is commonly accepted that both genes and cancer samples should be regarded as similar if they exhibit similarity in shape (trend similarity), rather than in the absolute differences in their values [4], [5], [1], [2], [31], [32]. Although related, these two

- *P.A. Jaskowiak and R.J.G.B. Campello are with the Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos, SP, Brazil. E-mail: {pablo, campello}@icmc.usp.br.*
- *I.G. Costa is with the Center of Informatics, Federal University of Pernambuco, Recife, PE, Brazil, and IZKF Computational Biology Research Group, Institute for Biomedical Engineering, Aachen University Medical School, RWTH Aachen, Germany. E-mail: igcf@cin.ufpe.br.*

*Manuscript received 4 Sept. 2012; revised 28 Jan. 2013; accepted 30 Jan. 2013; published online 11 Feb. 2013.*
*For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBBSI-2012-09-0236.*
*Digital Object Identifier no. 10.1109/TCBB.2013.9.*

[1] We distinguish the terms distance and distance metric as in [26].

applications, however, have their own peculiarities, as we discuss in the sequel.

For the clustering of genes (time-course experiments), there exist thousands of objects and only a dozen of features available. In fact, about 80 percent of all gene expression time-series have fewer than eight time points [12]. In this scenario, different experiments have distinct sampling frequencies and time resolutions; therefore, considering temporal dependencies between subsequent time points is crucial [12], [33]. Considering those aspects, several authors have proposed proximity measures specifically for the clustering of gene expression time-course data. The Jackknife correlation [4] was proposed to reduce the effect of single outliers on the final correlation value. The Short Time-Series Distance [34] explicitly takes into account differences among time intervals. The measure called Local Shape-Based Similarity [31] allows the detection of local, and possibly shifted, gene correlations. Finally, two measures that combine slope and range information with time-series correlation were presented in [32]. Although the last three proximity measures mentioned consider temporal dependencies between time points (feature order is important) this is not the case of *standard* measures, such as Pearson, which provides the same values regardless of the feature ordering.

For the clustering of cancer samples, one usually has a small number of objects embedded in a high-dimensional feature space composed of thousands of genes. Moreover, one assumes that the features (genes) are independent. Pearson [35] has been widely adopted as the rule of thumb in this scenario [36], [5], [2], [37], [38], even though, it is not the only alternative available to identify trend similarities. Finally, it has its own issues, e.g., it is not robust to outliers and noise [4], [2].

Despite of the wide variety of proximity measures available in the literature and of their key role in clustering, little attention has been given to the subject of proximity measures in gene expression data, as pointed out by the authors of [36], [39], [38]. Indeed, only a few works tried to provide guidelines concerning their choice for this particular domain [40], [41], [24], [25],[42], [43]. In [24] and [25], the authors reckon the importance of considering different proximity measures during the clustering of cancer samples, but they primarily focus on the comparison of clustering algorithms. In [40] and [41] proximity measures and clustering algorithms are compared for the clustering of time-course data, but measures specifically designed for this scenario are not taken into account. In [42] and [43], the authors compare, without any distinction, proximity measures regarding the clustering of both cancer samples and time-course data, despite the differences of such scenarios. Apart from [24], all other studies consider only a few data sets in their evaluations.

Regarding the methodologies that can be employed to evaluate proximity measures, we call attention to [42] and [43], in which the authors introduce the concept of *Intrinsic Separation Ability* (ISA). ISA can be employed to measure the agreement of a proximity measure with respect to a desired solution, i.e., a gold standard partition, *without* the need of a clustering algorithm. Although ISA provides information about the discriminative power of a distance, it can be applied solely to labeled data. Thereby, the use of ISA is restricted to cancer data sets, for which class labels are available. Given that labeled data are quite rare for gene clustering and time-course experiments, we introduce the concept of *Intrinsic Biological Separation Ability* (IBSA), which can measure the agreement of a proximity measure with respect to biological knowledge extracted from the Gene Ontology (GO) [44].

As a result of the lack of information concerning proximity measures for gene expression data clustering, Pearson is still the *de facto* proximity measure employed in this particular domain, whereas benefits that may arise from the adoption of other proximity measures remain uncertain. To the best of our knowledge, this is the first comparison specifically designed to evaluate proximity measures for gene expression data over a large number of real data sets. We evaluate proximity measures independently of biases of clustering algorithms, because both ISA and IBSA allow the evaluation of proximity measures without any clustering algorithm. We extend our previous work [45] in many aspects. Our main contributions can be summarized as follows:

- We compare proximity measures for the clustering of gene time-course data and cancer samples *separately*, as both scenarios have distinct characteristics.
- We evaluate a total of 16 proximity measures. For cancer data 10 measures are taken into account (we omit six measures that are time-course specific). For time-course data, all 16 measures are evaluated.
- For cancer samples, proximity measures are evaluated with respect to their ISA [42], as class labels are available.
- Based on [42] and [43], we introduce a new methodology to evaluate proximity measures for the clustering of genes, called IBSA. IBSA employs external information extracted from the GO [44] to overcome the lack of class labels in these data sets.
- Measures are evaluated regarding their robustness to noise, considering data with different noise levels.
- Our evaluation is performed in a substantial number of real data sets—52 data sets in total. From these, 35 come from a benchmark regarding cancer samples [24]. The other 17 are time-series data sets from the microarray literature, which we have preprocessed and compiled in a benchmark set. These data sets are publicly available at http://www.icmc.usp.br/~campello/Sub_Pages/IEEEACM_TCBB_arquivos/.

The remainder of the paper is organized as follows: In Section 2, we review the proximity measures considered in our work. In Section 3, we provide details regarding the evaluation of proximity measures, whereas in Section 4 we describe our experimental setup. The results of our evaluation are provided in Section 5. Finally, in Section 6, we address the main conclusions of our work.

## 2 Proximity Measures

Any two objects, genes or cancer samples, can be seen as real-valued sequences $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{b} = (b_1, \ldots, b_n)$, for which $n$ is the number of features. Having made such a consideration, we review the 16 proximity measures of our evaluation. First, we describe six correlation coefficients. Then, we review four "classical" proximity measures from the literature. Finally, we discuss in detail six measures specifically proposed by the authors of [4], [31], [34], and [32] for gene time-course data clustering.

## 2.1 Correlation Coefficients

Considering gene expression data, two objects (genes or samples) are usually regarded as similar if they exhibit similarity in shape (trend), rather than in absolute differences from their values. Therefore, correlation coefficients have been widely used, as they capture such a type of similarity. These measures provide values between $-1$ and 1. During our evaluation, each correlation was adapted to a distance in the following form:

$$\text{distance}(\mathbf{a}, \mathbf{b}) = 1 - \text{correlation coefficient}(\mathbf{a}, \mathbf{b}).$$

The equation above defines a distance but *not* a distance metric, since the triangle inequality is *not* satisfied [26].

### 2.1.1 Pearson

The Pearson correlation coefficient (PE) [35] allows the identification of linear correlations between sequences. It is described in (1), where $\bar{a}$ and $\bar{b}$ stand for the mean of sequences $\mathbf{a}$ and $\mathbf{b}$, respectively. Pearson may be sensitive to the presence of outliers, thus producing false positives, i.e., sequence pairs that are not alike, but receive a high correlation value [4], [2]. The PE has $O(n)$ time complexity

$$PE(\mathbf{a}, \mathbf{b}) = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2}\sqrt{\sum_{i=1}^{n}(b_i - \bar{b})^2}}. \quad (1)$$

### 2.1.2 Goodman-Kruskal

Goodman-Kruskal (GK) [46] takes into account only the ranks of $\mathbf{a}$ and $\mathbf{b}$. It is defined according to the number of concordant ($S_+$), discordant ($S_-$), and neutral pairs of elements in the sequences. In a concordant pair, the same relative order applies to both sequences, i.e., $a_i < a_j$ and $b_i < b_j$ or $a_i > a_j$ and $b_i > b_j$. For discordant pairs, the inverse relative order applies, i.e., $a_i < a_j$ and $b_i > b_j$ or $a_i > a_j$ and $b_i < b_j$. All other pairs are deemed neutrals. Goodman-Kruskal is given by (2). Note that it does not take into account neutrals in its normalization term and can reach its extrema even in their presence. Its time complexity is $O(n \; log \; n)$ [47]

$$GK(\mathbf{a}, \mathbf{b}) = \frac{S_+ - S_-}{S_+ + S_-}. \quad (2)$$

### 2.1.3 Kendall

The Kendall correlation coefficient (KE) [48] is based on the same building blocks used by Goodman-Kruskal. Kendall is defined in (3), where $n(n-1)/2$ is the total number of pairs of elements in the sequences. Note that, differently from GK, extreme correlation values are obtained only in the absence of neutrals. Kendall can be computed in $O(n \; log \; n)$ time [47]

$$KE(\mathbf{a}, \mathbf{b}) = \frac{S_+ - S_-}{n(n-1)/2}. \quad (3)$$

### 2.1.4 Spearman

The Spearman correlation (SP) [49] can be seen as a particular case of Pearson, provided that values of both $\mathbf{a}$ and $\mathbf{b}$ are replaced with their ranks in the respective sequences. Therefore, SP can also be defined by (1). As only the ranks of the sequences are considered, SP is more robust to outliers than Pearson [2]. SP has also been employed to gene expression data [6], [2], though less often than Pearson. Its time complexity is $O(n \; log \; n)$.

### 2.1.5 Rank Magnitude

The Rank-Magnitude correlation coefficient (RM) [47] was proposed as an asymmetric measure, for cases in which one of the sequences is composed of ranks and the other is composed by real numbers. It is defined by (4), with $R(a_i)$ denoting the rank of the $i^{th}$ position for sequence $\mathbf{a}$. In (4), $r^{min} = \sum_{i=1}^{n}(n+1-i)\bar{b}_i$ and $r^{max} = \sum_{i=1}^{n} i\bar{b}_i$. Value $\bar{b}_i$ corresponds to the $i$th element of the sequence obtained by rearranging sequence $\mathbf{b}$ so that it is sorted in ascending order

$$\hat{r}(\mathbf{a}, \mathbf{b}) = \frac{2\sum_{i=1}^{n} R(a_i)b_i - r^{max} - r^{min}}{r^{max} - r^{min}}. \quad (4)$$

We employ a symmetric version of RM [45], given by $RM(\mathbf{a}, \mathbf{b}) = (\hat{r}(\mathbf{a}, \mathbf{b}) + \hat{r}(\mathbf{b}, \mathbf{a}))/2$. This version can be used to compare two real-valued sequences, taking both their magnitudes and ranks into consideration. Any mention of RM in the remainder of this paper refers to its symmetric version, whose time complexity is $O(n \; log \; n)$.

### 2.1.6 Weighted Goodman-Kruskal

The Weighted Goodman-Kruskal correlation coefficient (WGK) [47] is presented in (5) and takes into consideration ranks and magnitudes of both sequences. Term $\hat{w}_{ij}$ is defined by (6). Terms $\hat{w}_{ij}^{\mathbf{a}}$ and $\hat{w}_{ij}^{\mathbf{b}}$ in (6) are defined by (7) and represent the signed percentage differences between the values of the $i$th and $j$th elements of the corresponding sequences. Term $w_{ij}$ in (5) is defined by (8), with $w_{ij}^{\mathbf{a}} = sign(a_i - a_j)$ and $w_{ij}^{\mathbf{b}} = sign(b_i - b_j)$. WGK correlation coefficient has $O(n^2)$ time complexity

$$WGK(\mathbf{a}, \mathbf{b}) = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \hat{w}_{ij}}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} |w_{ij}|} \quad (5)$$

$$\hat{w}_{ij} = \begin{cases} min\left\{\dfrac{\hat{w}_{ij}^{\mathbf{a}}}{\hat{w}_{ij}^{\mathbf{b}}}, \dfrac{\hat{w}_{ij}^{\mathbf{b}}}{\hat{w}_{ij}^{\mathbf{a}}}\right\} & \text{if } \hat{w}_{ij}^{\mathbf{a}} \quad \hat{w}_{ij}^{\mathbf{b}} > 0 \\ max\left\{\dfrac{\hat{w}_{ij}^{\mathbf{a}}}{\hat{w}_{ij}^{\mathbf{b}}}, \dfrac{\hat{w}_{ij}^{\mathbf{b}}}{\hat{w}_{ij}^{\mathbf{a}}}\right\} & \text{if } \hat{w}_{ij}^{\mathbf{a}} \quad \hat{w}_{ij}^{\mathbf{b}} < 0 \\ 1 & \text{if } \hat{w}_{ij}^{\mathbf{a}} = \hat{w}_{ij}^{\mathbf{b}} = 0 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

$$\hat{w}_{ij}^{\mathbf{x}} = \begin{cases} \dfrac{x_i - x_j}{x_{max} - x_{min}} & \text{if } x_{max} \neq x_{min} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

$$w_{ij} = \begin{cases} w_{ij}^{\mathbf{a}}/w_{ij}^{\mathbf{b}} & \text{if } w_{ij}^{\mathbf{b}} \neq 0 \\ 1 & \text{if } w_{ij}^{\mathbf{a}} = 0 \text{ and } w_{ij}^{\mathbf{b}} = 0 \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

## 2.2 Classical Measures

We review in the sequel four "classical" proximity measures that are also considered in our analysis. We anticipate that these measures have $O(n)$ time complexity.

### 2.2.1 Cosine Distance

The cosine similarity [5] is defined by (9) and can be regarded as the normalized inner product between $\mathbf{a}$ and $\mathbf{b}$. Note that the cosine similarity is related to Pearson and is sometimes referred to as uncentered correlation or angular separation [5]. The cosine measures the angle between two data points with respect to the origin, whereas Pearson correlation measures this angle considering the mean of the data

$$s_c(\mathbf{a}, \mathbf{b}) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} (a_i)^2}\sqrt{\sum_{i=1}^{n} (b_i)^2}}. \tag{9}$$

COS is straightforwardly obtained as $COS(\mathbf{a}, \mathbf{b}) = 1 - s_c(\mathbf{a}, \mathbf{b})$.

### 2.2.2 Minkowski Distance

One of the most popular proximity indices that measures dissimilarity between two data points is the Minkowski distance metric [19], defined by (10)

$$d_p(\mathbf{a}, \mathbf{b}) = \left( \sum_{i=1}^{n} |a_i - b_i|^p \right)^{1/p}. \tag{10}$$

Note that the Minkowski dissimilarity is parametric, i.e., for different values of parameter $p$ different dissimilarity measures are obtained. In our experiments, we considered three realizations of the Minkowski dissimilarity measure (three different values of the parameter $p$), which are the most commonly employed ones. These measures are commonly known as Manhattan distance (MAN) for $p = 1$, euclidean distance (EUC) for $p = 2$, and Supreme distance[2] (SUP) for $p = \infty$.

## 2.3 Time-Course Specific Measures

We review proximity measures specifically proposed for the clustering of gene time-course experiments. For these measures, we define $\mathbf{t} = (t_1, \dots, t_n)$ as the time instants at which each feature is measured for a gene.

### 2.3.1 Jackknife

The underlying idea behind the Jackknife (JK) correlation [4] is to minimize the effect of single outliers on the final correlation value by removing one single element at a time from both sequences. If the sequences do not contain outliers, their correlation value remains stable, otherwise, their removal causes a decrease in their correlation, indicating that the sequences were correlated partly due to the presence of outliers. JK is given by (11), where $PE^i(\mathbf{a}, \mathbf{b})$ is the Pearson correlation between $\mathbf{a}$ and $\mathbf{b}$ with their $i$th values removed. JK time complexity is $O(n^2)$, making its application to the clustering of cancer samples almost impractical

$$JK(\mathbf{a}, \mathbf{b}) = min\{PE^1(\mathbf{a}, \mathbf{b}), \dots, PE^n(\mathbf{a}, \mathbf{b}), PE(\mathbf{a}, \mathbf{b})\}. \tag{11}$$

### 2.3.2 Short Time-Series Dissimilarity

The Short Time-Series Dissimilarity (STS) was proposed in [34] and measures the distance between the $n - 1$ slopes that compound two gene time-series. For two genes $\mathbf{a}$ and $\mathbf{b}$, STS is given by (12). The greater the interval between the measurements, the lower its impact on the dissimilarity. Its time complexity is $O(n)$

$$STS(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^{n-1} \left( \frac{b_{i+1} - b_i}{t_{i+1} - t_i} - \frac{a_{i+1} - a_i}{t_{i+1} - t_i} \right)^2}. \tag{12}$$

### 2.3.3 Local Shape-Based Similarity

Based on the observation that biological relationships between genes may be present in the form of local and possibly shifted similarity patterns, [31] introduced the concept of Local Shape-based Similarity (LSS). LSS seeks the most similar subsequences of size $k$ in sequences $\mathbf{a}$ and $\mathbf{b}$. It is defined by (13) and (14), where $S$ is the base similarity employed between subsequences of a given size $k$. The minimum subsequence size is given by $k_{min}$, which is usually set to $n - 2$, allowing for two time instant shifts [31]. Note that although subsequences must have the same sizes, they do not have to be aligned, thus allowing locally shifted similarity patterns

$$LSS(\mathbf{a}, \mathbf{b}) = \max_{k_{min} \le k \le n} SIM_k(\mathbf{a}, \mathbf{b}) \tag{13}$$

$$SIM_k(\mathbf{a}, \mathbf{b}) = \max_{1 \le i,j \le n-k+1} S(\mathbf{a}[i, i + k - 1], \mathbf{b}[j, j + k - 1]). \tag{14}$$

In (13), LSS is given by the maximum similarity among sequences of different sizes. Since for shorter sequences greater similarities are more likely, LSS is derived from the probability of obtaining a specific similarity value for a given subsequence size. Similarity $S$ is given by the probability of obtaining a value of SP for the subsequences under comparison, as further detailed in [31]. LSS time complexity is $O(n^3)$, but it can be reduced by the use of approximations [31].

### 2.3.4 YR1 and YS1 Dissimilarities

Based on the presumption that correlations may not capture all information contained in gene time series, [32] introduced two dissimilarities that combine different types of information along with correlation values.

The first information taken into account by the authors concerns the agreement between the $n - 1$ slopes that compound the two gene expression series of size $n$ under comparison, as given by (15). Function $L$ is given by (16). Function $\mathcal{I}$ returns 1 in case of agreement and 0 otherwise. The slope of each gene $\mathbf{x}$, for a given interval, is given by (17)

$$A(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{n-1} \frac{\mathcal{I}(L(\mathbf{a}, i) = L(\mathbf{b}, i))}{n - 1} \tag{15}$$

$$L(\mathbf{x}, i) = \begin{cases} 1 & \text{if } slope(\mathbf{x}, i) > 0 \\ -1 & \text{if } slope(\mathbf{x}, i) < 0 \\ 0 & \text{if } slope(\mathbf{x}, i) = 0 \end{cases} \tag{16}$$

[2] SUP is sometimes referred to as Chebyshev distance. Given two objects, it corresponds to the greatest of their differences along any feature, i.e., the maximum difference among their features.

$$slope(\mathbf{x}, i) = \frac{x_{i+1} - x_i}{t_{i+1} - t_i}. \tag{17}$$

The second information concerns the agreement of maximum ($t^{max}$) and minimum ($t^{min}$) expression levels of both genes ($\mathbf{a}$ and $\mathbf{b}$), as given by

$$M(\mathbf{a}, \mathbf{b}) = \begin{cases} 1 & \text{if } t^{min}_{\mathbf{a}} = t^{min}_{\mathbf{b}} \text{ and } t^{max}_{\mathbf{a}} = t^{max}_{\mathbf{b}} \\ 0.5 & \text{if } t^{min}_{\mathbf{a}} = t^{min}_{\mathbf{b}} \text{ or } t^{max}_{\mathbf{a}} = t^{max}_{\mathbf{b}} \\ 0 & \text{if } t^{min}_{\mathbf{a}} \neq t^{min}_{\mathbf{b}} \text{ and } t^{max}_{\mathbf{a}} \neq t^{max}_{\mathbf{b}}. \end{cases} \tag{18}$$

The two proposed measures combine (15) and (18) with correlation coefficients, as given by (19) and (20). For $YR1$ and $YS1$, the authors also consider, respectively, PE and SP, in the forms: $R(\mathbf{a}, \mathbf{b}) = (PE(\mathbf{a}, \mathbf{b}) + 1)/2$ and $S(\mathbf{a}, \mathbf{b}) = (SP(\mathbf{a}, \mathbf{b}) + 1)/2$

$$YR1(\mathbf{a}, \mathbf{b}) = \omega_1 R(\mathbf{a}, \mathbf{b}) + \omega_2 A(\mathbf{a}, \mathbf{b}) + \omega_3 M(\mathbf{a}, \mathbf{b}) \tag{19}$$

$$YS1(\mathbf{a}, \mathbf{b}) = \omega_1 S(\mathbf{a}, \mathbf{b}) + \omega_2 A(\mathbf{a}, \mathbf{b}) + \omega_3 M(\mathbf{a}, \mathbf{b}). \tag{20}$$

In (19) and (20), the weight terms $\omega_1$, $\omega_2$, and $\omega_3$ must satisfy the condition $\sum_{i=1}^{3} \omega_i = 1$. Son and Baek [32] proposed a way to estimate their values, but their approach is highly computationally demanding. To compare the two measures, we set weights in the following form: $\omega_1 = 0.5$, $\omega_2 = 0.25$, and $\omega_3 = 0.25$, which are values employed by the authors in [32]. Following the suggestion from one of the reviewers, we also employed a measure called YRM1, based on Rank-Magnitude (Section 2.1.5), given by (21), where $\hat{R}(\mathbf{a}, \mathbf{b}) = (RM(\mathbf{a}, \mathbf{b}) + 1)/2$

$$YRM1(\mathbf{a}, \mathbf{b}) = \omega_1 \hat{R}(\mathbf{a}, \mathbf{b}) + \omega_2 A(\mathbf{a}, \mathbf{b}) + \omega_3 M(\mathbf{a}, \mathbf{b}). \tag{21}$$

$YR1$ has $O(n)$ computational time complexity while both $YS1$ and $YRM1$ have $O(n \ log \ n)$ time complexity.

## 3 Proximity Measure Evaluation

We are primarily interested in the effect that different proximity measures may have on the outcome of clustering problems. As different clustering algorithms provide different biases, it is virtually impossible to evaluate proximity measures considering all the different clustering algorithms proposed in the literature. Therefore, we chose to evaluate proximity measures based on their Intrinsic Separation Ability (ISA), i.e., the ability of each proximity measure to separate data points according to a previously given separation (external information) *without* the help of a particular clustering algorithm [42]. Our preliminary work [45] indicated a high correlation (agreement) between the ranking of proximity measures indicated by ISA and an evaluation of the measures conducted with different clustering algorithms. ISA was proposed by Giancarlo et al. [42], [43] for case in which class labels are available for a particular data set, i.e., when a gold standard partition (desired cluster memberships) is available.

ISA can only be computed for data sets with a gold standard partition, i.e., class labels. As class labels are usually unavailable for gene clustering (e.g., time-series data), we took advantage of the information provided by the Gene Ontology (GO) [44] to overcome the absence of labeled data and devise a new procedure to evaluate the ISA of a distance regarding the clustering of genes, called *Intrinsic Biological Separation Ability* (IBSA). In brief, IBSA assigns class labels according to semantic similarities among genes extracted from the GO (external biological information), which explains the term *biological* in its name. In the sequel, we describe ISA in detail and introduce our new methodology, IBSA.

### 3.1 Intrinsic Separation Ability

The ISA of a distance[3] indicates how well it can separate cancer samples without the influence of a clustering algorithm [42], [43]. Given a data set with $o$ objects (cancer samples) $\mathbf{x_1}, \ldots, \mathbf{x_o}$, we build a distance matrix $D$, where $D(i, j) = distance(\mathbf{x_i}, \mathbf{x_j})$, with $1 \leq i, j \leq o$. Assuming that all the values of $D$ are in the $[0, 1]$ interval (if they are not, they must be normalized), we proceed and build a binary classifier that assigns a pair of objects (cancer samples) to a given class according to (22), where $\phi_1$ is a given threshold in the $[0, 1]$ interval. By Applying (22) to all pairs of objects from a data set with a *fixed* threshold, we obtain a predicted solution, based solely on object distances

$$I_{\phi_1}(\mathbf{x_i}, \mathbf{x_j}) = \begin{cases} 1 & \text{if } D(i, j) \leq \phi_1 \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

Provided that we are dealing with labeled data in the case of cancer samples (below are details regarding data sets), we can proceed and build a desired solution for the classifier previously described. The desired solution is built upon the golden standard partition of each data set, given by (23) for all $\mathbf{x_i}$ and $\mathbf{x_j}$

$$J(\mathbf{x_i}, \mathbf{x_j}) = \begin{cases} 1 & \text{if } \mathbf{x_i} \text{ and } \mathbf{x_j} \text{ belong} \\ & \text{to the same cluster} \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

Note that by setting a threshold $\phi_1$ and applying (22) and (23) to all object pairs we have a predicted and a desired solution, respectively. For (22), however, the predicted solution is not unique, as different threshold values are possible. We consider all possible values[4] of $\phi_1$ in the $[0, 1]$ interval, generating a set of all possible predicted solutions for a given distance, i.e., one for each different value of $\phi_1$. To evaluate ISA for a given distance, we have to compare its predicted solutions against the expected one.

In brief, we have multiple comparisons to perform, one per each value of $\phi_1$. These comparisons can be addressed by employing the well-established concept of Receiver Operating Characteristics analysis, ROC analysis for short [50], [42], [43]. The whole procedure is as follows: First, given a threshold ($\phi_1$), we compute the number of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) given by

$$TP = \sum_{\forall i,j,i \neq j} I_{\phi_1}(\mathbf{x_i}, \mathbf{x_j}) J(\mathbf{x_i}, \mathbf{x_j})$$
$$FP = \sum_{\forall i,j,i \neq j} I_{\phi_1}(\mathbf{x_i}, \mathbf{x_j})(1 - J(\mathbf{x_i}, \mathbf{x_j}))$$

[3] As all the proximity measures employed in this paper are adapted as distances these terms will be used interchangeably from now on.

[4] The possible values of $\phi_1$ are those in the finite set of values contained in matrix $D$.

$$TN = \sum_{\forall i,j,i \neq j} (1 - I_{\phi_1}(\mathbf{x_i}, \mathbf{x_j}))(1 - J(\mathbf{x_i}, \mathbf{x_j}))$$

$$FN = \sum_{\forall i,j,i \neq j} (1 - I_{\phi_1}(\mathbf{x_i}, \mathbf{x_j}))J(\mathbf{x_i}, \mathbf{x_j}).$$

Then, we compute the False Positive Rate, given by $FPR = FP/(FP + TN)$, and the True Positive Rate, given by $TPR = TP/(TP + FN)$. Values of FPR and TPR are computed for all the values of $\phi_1$. With these values in hand, we plot an ROC Curve (FPR versus TPR), which is summarized by its Area Under the Curve (AUC). AUC values are in the $[0, 1]$ interval. In this particular case, an AUC value of 1 indicates a distance measure that perfectly separates cancer samples according to the desired solution. On the other hand, an AUC value close to or lower than 0.5 indicates a distance measure that fails to separate objects according to the desired solution. Finally, the ISA of a distance is given by its AUC.

### 3.2 Intrinsic Biological Separation Ability

The ISA can be computed only for data sets with a golden standard partition, i.e., data sets for which class labels are available. Note that for most gene clustering problems, as time-series data sets, no class labels are available. Therefore, we take advantage of the information provided by the GO [44] to overcome the lack of labeled data and devise a new procedure to evaluate the ISA of a distance regarding the clustering of genes. This new procedure is called *Intrinsic Biological Separation Ability* (IBSA). Instead of using class labels, our methodology employs external biological information (semantic similarities among genes) extracted from the GO. Note that since IBSA employs information from the GO to evaluate a particular proximity measure, it tends to favor proximity measures that are in agreement with GO external information. If the user is interested in finding a different type of structure in the data (not related with GO), another methodology should be selected and employed.

Given a data set with $o$ objects (genes) $\mathbf{x_1}, \ldots, \mathbf{x_o}$ we build a distance matrix $D$, where $D(i,j) = distance(\mathbf{x_i}, \mathbf{x_j})$, with $1 \leq i, j \leq o$. Assuming that all the values of $D$ are in the $[0, 1]$ interval, all pairs of objects can be distinguished by the same binary classifier previously described by (22). In brief, object pairs are assigned to class 1 if the distance between them is smaller than or equal to a given threshold $\phi_1$ in the $[0, 1]$ interval and 0 otherwise. Applying this equation to all object pairs from a given data set (with a *fixed* threshold), we obtain a predicted solution based solely on the distances between object pairs.

To build a desired solution for this classifier, the first step of our methodology consists in obtaining biological dissimilarities for all *pairs of genes* from the data set in hand, devising a biological dissimilarity matrix ($B$). Considering the GO, several proximity measures can be employed to quantify the degree of concordance between the *sets of terms* that annotate any two genes. By combining dissimilarities that operate between *sets of terms*, it is possible to measure the degree of concordance between any two genes [51]. Note that the methodology presented here is the same regardless of the biological similarity employed between genes. Therefore, we elaborate on the choice of the biological measure during the discussion of the Experimental Setup (Section 4).

Once a biological dissimilarity matrix is available, it can be interpreted as external information and fill the gap left by the lack of class labels. For a given biological dissimilarity matrix ($B$) with values in the $[0, 1]$ interval, we proceed and build a desired biological solution, as specified by (24), where $\phi_2$ is a threshold in the $[0, 1]$ interval. By applying (24) to all pairs of objects from a given data set (with a *fixed* threshold), we obtain a desired biological solution, based on external information extracted from the GO

$$J_{\phi_2}(\mathbf{x_i}, \mathbf{x_j}) = \begin{cases} 1 & \text{if } B(i,j) \leq \phi_2 \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

Note that by fixing thresholds $\phi_1$ and $\phi_2$ and applying (22) and (24) to all object pairs we have a predicted and a desired biological solution, respectively. As we are dealing with two thresholds, there exist two sets of solutions: 1) a set of predicted solutions, obtained when (22) is applied to all pairs of objects with $\phi_1$ taking all its possible values, and 2) a set of desired biological solutions, obtained by applying (24) to all pairs of objects with all possible values for $\phi_2$. All the possible values for $\phi_1$ and $\phi_2$ are the values of the elements in matrices $D$ and $B$, respectively. To evaluate the IBSA of a given distance, its set of predicted solutions must be compared against the set of expected ones, obtained from the GO.

Since two thresholds are employed in our methodology, multiple ROC analyses must be performed for the different values of $\phi_1$ and $\phi_2$. First, we fix a value for $\phi_2$ obtaining a desired biological solution. With a desired solution in hand, the analysis is performed as previously described for ISA, i.e.: 1) with a fixed biological solution, we set a value for $\phi_1$ and obtain the values of TP, FP, TN, and FN, which are given by

$$TP = \sum_{\forall i,j,i \neq j} I_{\phi_1}(\mathbf{x_i}, \mathbf{x_j})J_{\phi_2}(\mathbf{x_i}, \mathbf{x_j})$$

$$FP = \sum_{\forall i,j,i \neq j} I_{\phi_1}(\mathbf{x_i}, \mathbf{x_j})(1 - J_{\phi_2}(\mathbf{x_i}, \mathbf{x_j}))$$

$$TN = \sum_{\forall i,j,i \neq j} (1 - I_{\phi_1}(\mathbf{x_i}, \mathbf{x_j}))(1 - J_{\phi_2}(\mathbf{x_i}, \mathbf{x_j}))$$

$$FN = \sum_{\forall i,j,i \neq j} (1 - I_{\phi_1}(\mathbf{x_i}, \mathbf{x_j}))J_{\phi_2}(\mathbf{x_i}, \mathbf{x_j}).$$

Then, we compute the False Positive Rate ($FPR = FP / (FP + TN)$) and the True Positive Rate ($TPR = TP / (TP + FN)$) for different values of $\phi_1$; 2) we plot an ROC Curve and obtain its corresponding AUC value. This procedure is repeated for all different values of $\phi_2$, i.e., the set of desired biological solutions. Distances are evaluated by the *sum* of their AUCs, which we normalize in the $[0, 1]$ interval for better interpretation. Normalization is given by the average of the sum of the AUCs.

By applying such a methodology, we can verify whether the IBSA of a particular distance is in agreement with the biological distance extracted from the GO. Note that this evaluation method also avoids the bias of employing a particular clustering algorithm. We refer to this novel methodology as the *Intrinsic Biological Separation Ability* of a proximity measure (distance), IBSA for short.

## 4 Experimental Setup

### 4.1 Cancer Data

For the evaluation regarding cancer experiments, we considered the 35 benchmark data sets introduced in [24].

TABLE 1
Cancer Data Sets Used in the Experiments

| | Name | $\#s$ | $\#c$ | $\#g$ |
|---|---|---|---|---|
| cDNA | *alizadeh-v1* | 42 | 2 | 1095 |
| | *alizadeh-v2* | 62 | 3 | 2093 |
| | *alizadeh-v3* | 62 | 4 | 2093 |
| | *bittner* | 38 | 2 | 2201 |
| | *bredel* | 50 | 3 | 1739 |
| | *chen* | 180 | 2 | 85 |
| | *garber* | 66 | 4 | 4553 |
| | *khan* | 83 | 4 | 1069 |
| | *lapointe-v1* | 69 | 3 | 1625 |
| | *lapointe-v2* | 110 | 4 | 2496 |
| | *liang* | 37 | 3 | 1411 |
| | *risinger* | 42 | 4 | 1771 |
| | *tomlins-v1* | 104 | 5 | 2315 |
| | *tomlins-v2* | 92 | 4 | 1288 |
| Affymetrix | *armstrong-v1* | 72 | 2 | 1081 |
| | *armstrong-v2* | 72 | 3 | 2194 |
| | *bhattacharjee* | 203 | 5 | 1543 |
| | *chowdary* | 104 | 2 | 182 |
| | *dyrskjot* | 40 | 3 | 1203 |
| | *golub-v1* | 72 | 2 | 1877 |
| | *golub-v2* | 72 | 3 | 1877 |
| | *gordon* | 181 | 2 | 1626 |
| | *laiho* | 37 | 2 | 2202 |
| | *nutt-v1* | 50 | 4 | 1377 |
| | *nutt-v2* | 28 | 2 | 1070 |
| | *nutt-v3* | 22 | 2 | 1152 |
| | *pomeroy-v1* | 34 | 2 | 857 |
| | *pomeroy-v2* | 42 | 5 | 1379 |
| | *ramaswamy* | 190 | 14 | 1363 |
| | *shipp* | 77 | 2 | 798 |
| | *singh* | 102 | 2 | 339 |
| | *su* | 174 | 10 | 1571 |
| | *west* | 49 | 2 | 1198 |
| | *yeoh-v1* | 248 | 2 | 2526 |
| | *yeoh-v2* | 248 | 6 | 2526 |

TABLE 2
Time-Course Data Sets Used in the Experiments

| Name | Source | $\#s$ | $\#\hat{g}$ | $\#g$ |
|---|---|---|---|---|
| *alpha factor* | Spellman *et al.* (1998) | 18 | 6178 | 1099 |
| *cdc 15* | | 24 | 6178 | 1086 |
| *cdc 28* | | 17 | 6178 | 1044 |
| *elutriation* | | 14 | 6178 | 935 |
| *1mM menadione* | Gasch *et al.* (2000) | 9 | 6152 | 1050 |
| *1M sorbitol* | | 7 | 6152 | 1030 |
| *1.5mM diamide* | | 8 | 6152 | 1038 |
| *2.5mM DTT* | | 8 | 6152 | 991 |
| *constant 32nM H2O2* | | 10 | 6152 | 976 |
| *diauxic shift* | | 7 | 6152 | 1016 |
| *complete DTT* | | 7 | 6152 | 962 |
| *heat shock 1* | | 8 | 6152 | 988 |
| *heat shock 2* | | 7 | 6152 | 999 |
| *nitrogen depletion* | | 10 | 6152 | 1011 |
| *YPD 1* | | 12 | 6152 | 1011 |
| *YPD 2* | | 10 | 6152 | 1022 |
| *yeast sporulation* | Chu *et al.* (1998) | 7 | 6118 | 1171 |

In brief, this publicly available benchmark collection encompasses 35 microarray data sets from cancer gene expression experiments and comprehend the two platforms in which the technology is generally available, i.e., Affymetrix (21 data sets) and cDNA (14 data sets) [2]. All data sets are already preprocessed, with the most significant preprocessing performed by Souto et al. [24] being related to the removal of uninformative genes (genes that are not differentially expressed across samples). A summary of the 35 cancer data sets is shown in Table 1, where $\#s$, $\#g$, and $\#c$ correspond to the number of cancer samples (objects), genes (features), and prelabeled clusters of each data set. Information about all data sets and details of preprocessing can be obtained in [24].

Since for cancer data we are dealing with labeled data, we have adopted ISA [42], as discussed in Section 3.1.

### 4.2 Time-Course Data

For the clustering of genes, we considered time-course data sets from three different sources (the three sources are shown in Table 2). These data sets are from cDNA microarrays experiments regarding the *Saccharomyces cerevisiae* organism (yeast). Two data sources comprehend multiple time-series experiments and are further divided into single experiment data sets. Overall there is a total of 17 data sets. Before performing the experiments, we removed genes for which 10 percent or more expression values are missing. After this removal no gene with missing values remained. Finally, for further analysis, we selected genes that displayed a difference of at least $l$-fold in at least $c$ samples from their mean expression level [52]. We considered $c = 1$ and adjusted the value of $l$ to select about 1,000 genes, number employed in several studies (e.g., [53], [54], [55], [24], ). A summary of the 17 data sets is shown in Table 2, where $\#s$, $\#\hat{g}$, and $\#g$ denote the number of time points (samples), original number of genes for each data set, and number of genes obtained after the filtering procedure, respectively. A compiled version of these 17 preprocessed data sets is available as a benchmark set online at http://www.icmc.usp.br/~campello/Sub_Pages/IEEEACM_TCBB_arquivos/.

Following the suggestion from one of the reviewers, we also applied normalization procedures to data sets from [56], [57], [54]. We employed the procedure called Multiple-Slide Normalization, as described in [58]. Due to space constraints, information and results concerning the normalized data sets are provided in our supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2013.9. We anticipate, however, that the conclusions of our work are basically the same for original and normalized data, as they provided similar results.

Since we are dealing with unlabeled data in the case of time-course data sets, we employed the IBSA analysis (Section 3.2) to evaluate proximity measures. IBSA requires the definition of a biological proximity measure between genes to generate a biological dissimilarity matrix for each data set. We employed the measure proposed by Resnik [59], as previous work has shown that it correlates best with both gene expression similarity patterns [60] and gene sequence pattern similarities [29], [51]. We used the Best-Match Average (BMA) of the Resnik measure, as it provides better results than other approaches employed to combine ontology term similarities [29].

Given two GO terms $\mathrm{t}_1$ and $\mathrm{t}_2$, Resnik's similarity is given by (25), where $S(\mathrm{t}_1, \mathrm{t}_2)$ is the set of terms that subsume both $\mathrm{t}_1$ and $\mathrm{t}_2$, i.e., common ancestors of both $\mathrm{t}_1$ and $\mathrm{t}_2$. For a given common ancestor $\mathrm{t}$, $p(\mathrm{t})$ is the probability of annotating a gene with it, whereas $[-\log p(\mathrm{t})]$ is usually referred to as the Information Content (IC) of term $\mathrm{t}$. For our experiments, $p(\mathrm{t})$ (probability values) were estimated with the GOSim R Package [61]. Such an estimation is based on empirical observations regarding the number of times that a GO term annotates a gene [61]. Resnik's similarity seeks for the common ancestor $\mathrm{t}$ with greatest IC. In fact, Resnik's similarity

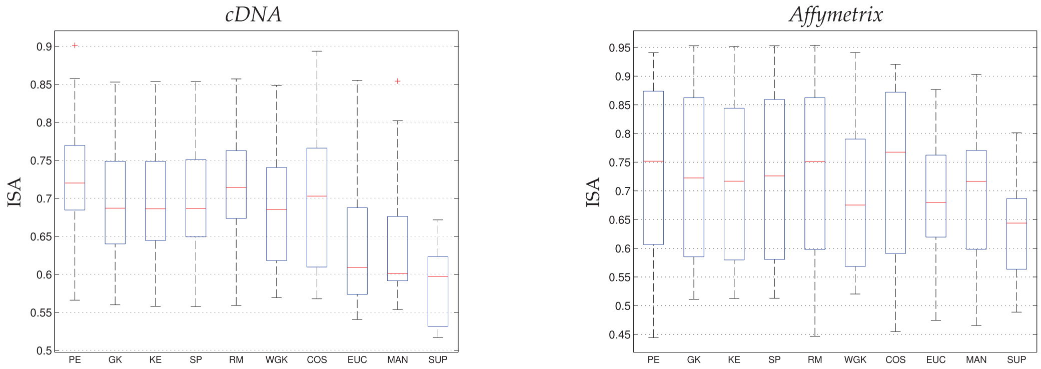

Fig. 1. Boxplots depict the ISA for each one of the evaluated distances.

between two terms is equal to the greatest IC among the term's common ancestors

$$\mathrm{Resnik}_{\mathrm{Sim}}(\mathrm{t}_1, \mathrm{t}_2) = \max_{\mathrm{t} \in S(\mathrm{t}_1, \mathrm{t}_2)} [-\log p(\mathrm{t})]. \tag{25}$$

Note that Resnik's similarity is computed only between two terms. As one gene may be annotated with a set of terms, we need to obtain gene similarities based on the sets of terms that annotate each gene under consideration. To obtain *gene similarities*, we combined term similarities, obtained with (25), employing the BMA [29] of the Resnik similarity, as given by (26). In this equation, $\mathcal{S}_1$ and $\mathcal{S}_2$ are the sets of GO terms that annotate genes $g_1$ and $g_2$, respectively. For use with IBSA, BMA dissimilarities can be easily obtained from $\mathrm{BMA}_{\mathrm{Dis}} = 1 - \mathrm{BMA}_{\mathrm{Sim}}$

$$\begin{aligned}\mathrm{BMA}_{\mathrm{Sim}}(g_1, g_2) = &\frac{1}{2}\mathrm{avg}_{i \in \mathcal{S}_1}(\max{}_{j \in \mathcal{S}_2}\ \mathrm{Resnik}_{\mathrm{Sim}}(\mathrm{t}_i, \mathrm{t}_j)) \\ &+ \frac{1}{2}\mathrm{avg}_{j \in \mathcal{S}_2}(\max{}_{i \in \mathcal{S}_1}\ \mathrm{Resnik}_{\mathrm{Sim}}(\mathrm{t}_i, \mathrm{t}_j)).\end{aligned} \tag{26}$$

Finally, we computed IBSA for time-course data sets considering the Molecular Function (MF) and Biological Process (BP) ontologies separately. Cellular Component (CC) ontology was left out as it usually shows little correlation with gene expression data [62], [63].

### 4.3 Evaluating Robustness to Noise

To evaluate robustness to noise, we artificially added noise to some data sets (we selected eight data sets for such experiments, as we detail in the sequel). Since we did not know a priori the amount of noise originally present in each data set, as a first step, we chose eight data sets (four for cancer experiments and four for time-course experiments) in which differences among the performances of the distances were minimal, i.e., data sets in which different distances provided the closest results without the presence of noise. With this selection, we intended to provide a fair starting point and comparison among the distances as the noise is added.

After choosing appropriate data sets, we added noise to each data set as follows: 1) we randomly chose $\alpha$ percent *points* in the data set (we call *point* a specific expression level regarding one gene and one sample, i.e., a cell of the data matrix) and; 2) replaced each point with a randomly generated value between the minimum value and the maximum value observed in the data matrix. This procedure was performed for values of $\alpha$ ranging from 1 to 10 percent, considering an increment step of 1 percent. To increase the reliability of the results, 100 different noisy data sets were generated for each different percentage of noise (as indicated in [19], at least 100 replications should be considered for obtaining statistical significance). The replications were required to avoid bias in the errors introduced in particular data sets. Such experiments were based on a usual Monte Carlo analysis proposed in [64]. Replications were sampled independently of each other.

### 4.4 Assessing Statistical Significance

To assure the validity of the results of our comparison, we employed Friedman and Nemenyi statistical tests [65]. These two tests are suitable when comparing multiple proximity measures across multiple data sets [65], which is the case in our study. The Friedman statistical test is a nonparametric counterpart of the well-known ANOVA, and is based on the ranking of the algorithms under comparison. If the null hypothesis of the Friedman statistical test[5] is rejected, then the Nemenyi statistical test is applied to identify among which pairs of algorithms the difference exists. For a review of Friedman and Nemenyi statistical tests, we refer the reader to [65].

## 5 Results and Discussion

In the sequel, we present results of the comparison of distances for both cancer and time-course experiments. Note that, due to space restrictions, for time-course data set we report results considering the data sets as provided by their original authors, i.e., without normalization. Results in agreement with the ones presented here, concerning the normalized versions of such data sets, can be found in our supplementary material, available online.

### 5.1 Cancer Tissue Experiments

We summarize in Fig. 1 results regarding the ISA of the distances, for each data set type. For cDNA data sets, PE provided the best results, followed by RM and COS. When considering Affymetrix data, RM, COS, and PE stood out. Note that, regardless of the type of data, rank-based measures, i.e., GK, KE, and SP, provided, along with

[5] That states that all algorithms under comparison are equivalent.

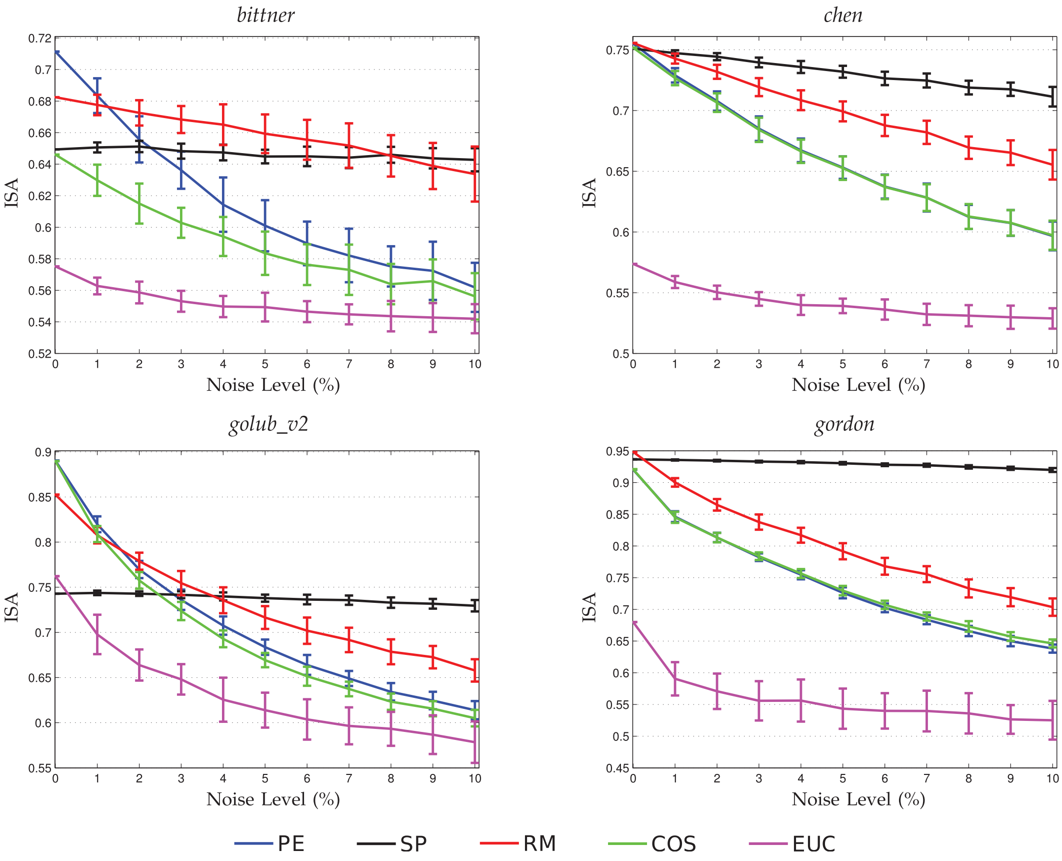

Fig. 2. ISA for different noise levels (%) regarding Pearson (PE), Spearman (SP), Rank-Magnitude (RM), Cosine (COS), and euclidean distance (EUC) in four data sets. Lines correspond to mean ISA values (bars account for standard deviations) for executions performed in 100 noisy data sets for each noise level between 1 and 10 percent.

WGK, worse results than the other correlation coefficients. This behavior can be interpreted as a consequence of the loss of information intrinsic to rank-based measures. It is worth noting, however, that measures, such as GK and KE, rarely employed in the gene expression literature, appear as good alternatives to the commonly employed SP. Except for COS, which is closely related to PE, all "classical" distance functions (EUC, MAN, and SUP) provided the worst overall ISAs.

We employed Friedman and Nemenyi statistical tests (at 95 percent confidence level) separately for cDNA and Affymetrix data. Regarding cDNA, the tests suggest that PE and RM provided better ISA than EUC, MAN, and SUP distances. Considering Affymetrix, RM and GK provided superior results than SUP. For complete results in tabular form, please refer to our supplementary file, available online.

Fig. 2 depicts results considering different levels of noise. Results concern four data sets, namely, *bittner* and *chen* from cDNA and *golub V2* and *gordon* from Affymetrix. For simplicity, we analyzed the three distances that produced superior or competitive results in the previous evaluation, i.e., PE, RM, and COS. We also included the commonly employed EUC and SP. We observed that KE and GK provided similar results when compared to SP. These two correlation coefficients were not included to keep the plots clearer.

According to the results from Fig. 2, SP stands out as the distance less affected by the presence of noise for both cDNA and Affymetrix data sets. These results are not surprising, because SP is a rank-based correlation and has been widely used due to its known robustness. As it takes into account only the ranks of the values for each sample compared, small disturbances in the data tend to vanish in the final comparison. Although RM is more sensitive to the presence of noise than SP, it provided better results than the other distances. PE and COS provided similar results among themselves, whereas the commonly employed EUC showed the worst results.

It is worth noting that, although SP provided the best results w.r.t. different levels of noise, it provided inferior mean results when compared to PE, RM, and COS in the previous evaluation scenarios (no noise added). To this extent, RM appears to be one of the best alternatives among the compared measures, as it: 1) figures among the top two measures in almost all cases in the previous evaluation scenario and 2) is more robust to noise than PE and COS. In brief, RM provides a good compromise between accuracy and robustness to noise.

## 5.2 Gene Time-Course Experiments

IBSA results are shown in Fig. 3, for MF and BP ontologies. The best results were obtained with YS1, which was specifically proposed to clustering gene expression time-course data. JK, another measure proposed specifically to this particular scenario, also provided slightly better results than PE. YR1, YRM1, COS, and WGK also stood out as good choices regarding IBSA. Note that YRM1 provided better

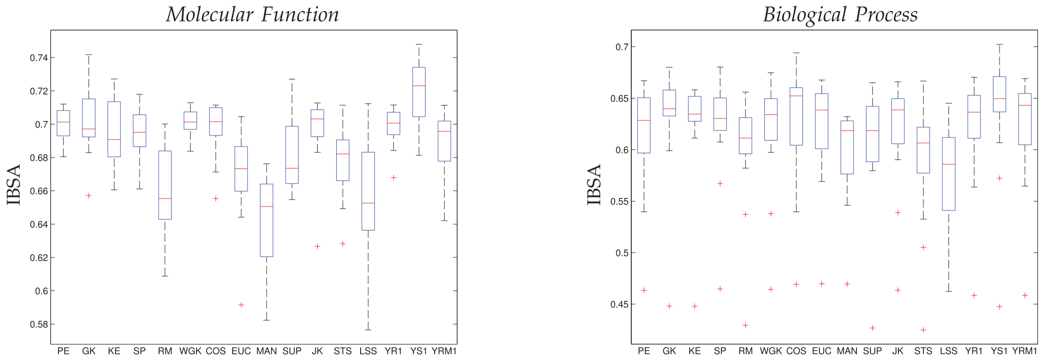

Fig. 3. Boxplots depict the IBSA (IBSA values) for each one of the evaluated distances.

results than its base measure (RM). Considering rank-based measures, the best results were obtained by GK, which provided in some cases even better results than PE and JK. Regarding the remaining distances, some trends can be observed: 1) classical distances (except COS) provided inferior results when compared to the other measures; 2) RM showed inferior results when compared to the remaining correlations; and 3) STS and LSS did not provide good IBSA values.

We employed Friedman and Nemenyi statistical tests (at 95 percent confidence level) separately for the results obtained with MF and BP ontologies. The results are shown in Table 3. Once again, PE and JK exhibited very similar results. Both measures provided statistically better results than EUC for the MF ontology. When considering both ontologies, YS1 provided a greater number of statistically significant differences than PE and JK with respect to other proximity measures.

Apart from the results observed with the original data sets, we depict in Fig. 4 results regarding IBSA of the distances when different levels of noise were considered. We selected four data sets for evaluation as discussed in Section 4.3, namely, *alpha factor* and *cdc 15*, for the MF and *1.5 mM diamide* and *YPD 1*, for the BP ontology. To provide clearer plots, we selected, based on previous experiments, seven proximity measures for this evaluation: PE, SP, GK, EUC, JK, YR1, and YS1.

TABLE 3
Statistical Test Summary—MF and BP Ontologies

| | PE | GK | KE | SP | RM | WGK | COS | EUC | MAN | SUP | JK | STS | LSS | YR1 | YS1 | YRM1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PE | — | | | | | | | | | | | | | | | |
| GK | | — | | | | | | | | | | | | | | |
| KE | | | — | | | | | | | | | | | | * | |
| SP | | | | — | | | | | | | | | | | | |
| RM | * | ⧆ | | □ | — | * | * | | | | * | | | | ⧆ | |
| WGK | | | | | | — | | | | | | | | | | |
| COS | | | | | | | — | | | | | | | | | |
| EUC | * | | | | | * | * | — | | | * | | | | * | |
| MAN | * | ⧆ | * | * | | * | * | | — | | * | | | * | ⧆ | * |
| SUP | | | | | | | | | | — | | | | | ⧆ | |
| JK | | | | | | | | | | | — | | | | | |
| STS | | □ | | □ | | | □ | | | | | — | | | ⧆ | □ |
| LSS | * | ⧆ | ⧆ | ⧆ | | ⧆ | ⧆ | □ | | | * | | — | ⧆ | ⧆ | □ |
| YR1 | | | | | | | | | | | | | | — | | |
| YS1 | | | | | | | | | | | | | | | — | |
| YRM1 | | | | | | | | | | | | | | | | — |

Symbols in each cell denote that the measure in the column outperformed the one in the row regarding: * MF ontology, □ BP ontology, ⧆ both ontologies.

Considering the results in Fig. 4, two distinct groups of proximity measures can be observed: One composed of rank-based measures and another formed by the remaining proximity measures. Rank-based proximity measures were less affected by noise, exhibiting slower reductions in their IBSA values when compared to other proximity measures. Note that JK and YR1 provided little or no improvement when compared to PE (their base measure). Within the two groups of measures, it is difficult to determine whether one measure provides better results than the others, given the similar trends among their declines in IBSA.

For this particular scenario, YS1, which provided one of the top results regarding the comparison performed in the original data sets, was also one of the least affected measures subjected to different levels of noise. Although SP also showed good to moderate robustness to noise, it provided worse results in comparison to YS1 in the original data sets. Therefore, YS1 stands out as one of the best measures regarding gene time-course data, providing not only accurate results, but also robustness to noise.

## 5.3 Discussion

### 5.3.1 Regarding Both Types of Data

Kendall and Goodman-Kruskal, which are rarely employed, appeared as possible alternatives to the commonly employed Spearman. Apart from Cosine, which produced quite competitive results in a number of cases, other "classical" distances, i.e., Manhattan and Supreme distances, should be avoided. Although euclidean distance is usually employed to gene expression data clustering, it was not robust to noise in our experiments. Moreover, our results support that other proximity measures provided better results, as we discuss separately for each scenario in the sequel. For these reasons, euclidean distance can be kept as an alternative, but not as a first choice.

It is worth noting that the results displayed some variability for both data types, i.e., different results were obtained for different data sets. Such a variability has also been noted in studies that employed a considerable number of gene expression data sets, such as [24] and [25]. As a consequence, in some data sets, a proximity measure that does not provide the best mean (or median) results may turn out to be the best choice.

### 5.3.2 Regarding Cancer Data Sets

PE, RM, and COS provided competitive results among each other and superior results when compared to other

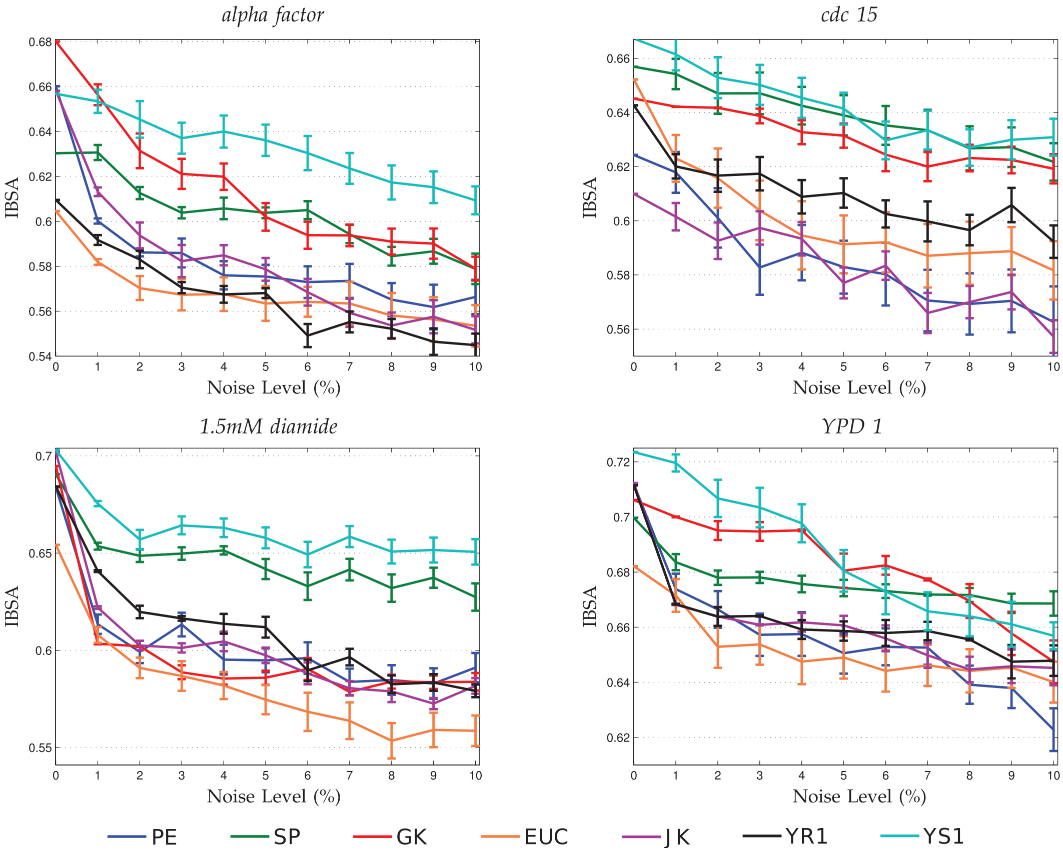

Fig. 4. IBSA for different noise levels (%) regarding Pearson (PE), Spearman (SP), Goodman-Kruskal (GK), euclidean distance (EUC), Jackknife (JK), YR1, and YS1 in four data sets. Lines correspond to mean IBSA values (bars account for standard deviations) for executions performed in 100 noisy data sets for each noise level between 1 and 10 percent.

measures. Rank-based measures, i.e., GK, Kendall, and SP led to inferior results in comparison to the aforementioned ones. This particular behavior may be explained by the loss of information inherent to their definition. These measures, however, have shown to be more robust when noise is added to the data. By combining sequence values along with their ranks, RM showed to be less sensitive to noise than PE, COS, and EUC (which employ solely sequence value information), though more sensitive to noise than Spearman, Kendall, and GK (which use only rank information). As RM has shown to be more accurate than rank-based measures, it emerges as a promising choice exhibiting a good compromise between accuracy and robustness to noise, with moderate time complexity.

### *5.3.3 Regarding Time-Course Data Sets*

In this scenario, YS1 showed to be one of the best proximity measures available. YS1 not only provided a good IBSA but also showed to be quite robust to noise. Such results were achieved by combining slope and range information with time-series correlation, giving rise to a measure that considers temporal dependencies between time points, i.e., their order. On the one hand, the additional information employed by the measure seems to improve accuracy; on the other hand the use of SP correlation as a base measure affords robustness to noise.

We observed that both YR1 and YS1 provided better results than their "base" proximity measures, i.e., PE and SP. We call attention to the fact that YR1 and YS1 are parametrized and the tuning of their parameters for a particular data set[6] may provide even better results than the ones observed here.

Regarding the other three proximity measures specifically proposed for the clustering of time-course data, we noted that JK provided little improvement in comparison to Pearson. If any, an improvement comes with the price of a quadratic time complexity, which, however, may not be an issue for *short* time course. LSS and STS figured among the worst proximity measures under evaluation. Although time shifts are important, considering them in series of limited size may not provide reliable information. Possibly for this reason, poor results were observed with LSS. As for STS, its simple formulation based solely on slope differences may be the root cause for the poor results observed. Note that when this information is combined with other, better results arise, as both YR1 and YS1 employ slope information as a *part* of their formulation.

Finally, for the remaining proximity measures, we observed that among rank-based measures (GK, SP, Kendall) the best results were obtained with GK. In this particular scenario, Weighted GK also provided good results. Finally, RM provided the worst results in comparison to the other correlation coefficients. We recall that RM

[6] As previously discussed in Section 2, we employed fixed parameters for these two proximity measures, given the large number of data sets and the computational time required in the experiments.

figured among the top proximity measures for cancer data sets. This difference in performance highlights that both scenarios posses quite different characteristics which should be always taken into account.

## 6 Conclusions

We have conducted a review and comparison of 16 proximity measures for the clustering of gene expression data. We considered six correlation coefficients, four "classical" distances, and six proximity measures specifically proposed for the clustering of gene time-course data. Given their differences, we evaluated proximity measures separately for cancer and time-course experiments. Apart from the comparison of proximity measures, we introduced a set of 17 time-course benchmark data along with a new methodology (IBSA) to evaluate distances for the clustering of genes. Both data sets and methodology can be used in future research to evaluate the effectiveness of new proximity measures in this particular scenario. IBSA can be employed to evaluate proximity measures regarding any gene clustering application, i.e., it is not restricted to gene time-course data, the scenario addressed here.

We highlight that cancer and time-course experiments posses quite different characteristics, which should be taken into account when selecting a proximity measure. For these two application scenarios, two different proximity measures stood out as promising alternatives, i.e., RM for cancer data and YS1 for time-course experiments. They show not only good performances, but also robustness in the presence of noise. Therefore, RM and YS1 should be considered as alternatives to more commonly employed measures and added to the toolbox of the field practitioners.

## Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the manuscript. The authors also acknowledge Brazilian research agencies CAPES, CNPq, FACEPE, FAPESP (process #2011/04247-5) and IZKF, RWTH University Hospital for financial support.

## References

[1] D. Jiang, C. Tang, and A. Zhang, "Cluster Analysis for Gene Expression Data: A Survey," *IEEE Trans. Knowledge Data Eng.*, vol. 16, no. 11, pp. 1370-1386, Nov. 2004.

[2] A. Zhang, *Advanced Analysis of Gene Expression Microarray Data*, first ed. World Scientific, 2006.

[3] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proc. Nat'l Academy Sciences USA*, vol. 95, no. 25, pp. 14863-14868, 1998.

[4] L.J. Heyer, S. Kruglyak, and S. Yooseph, "Exploring Expression Data: Identification and Analysis of Coexpressed Genes," *Genome Research*, vol. 9, no. 11, pp. 1106-1115, 1999.

[5] P. D'haeseleer, "How Does Gene Expression Clustering Work?" *Nature Biotechnology*, vol. 23, no. 12, pp. 1499-1501, 2005.

[6] G. Kerr, H.J. Ruskin, M. Crane, and P. Doolan, "Techniques for Clustering Gene Expression Data," *Computers Biology Medicine*, vol. 38, no. 3, pp. 283-293, 2008.

[7] T.R. Golub et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, pp. 531-537, 1999.

[8] A.A. Alizadeh et al., "Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling." *Nature*, vol. 403, no. 6769, pp. 503-511, 2000.

[9] S. Ramaswamy, K.N. Ross, E.S. Lander, and T.R. Golub, "A Molecular Signature of Metastasis in Primary Solid Tumors," *Nature Genetics*, vol. 33, no. 1, pp. 49-54, Jan. 2003.

[10] J. Lapointe et al., "Gene Expression Profiling Identifies Clinically Relevant Subtypes of Prostate Cancer," *Proc Nat'l Academy Sciences USA*, vol. 101, no. 3, pp. 811-816, 2004.

[11] I.G. Costa, A. Schönhuth, and A. Schliep, "The Graphical Query Language: A Tool for Analysis of Gene Expression Time-Courses," *Bioinformatics*, vol. 21, no. 10, pp. 2544-2545, 2005.

[12] J. Ernst, G.J. Nau, and Z. Bar-Joseph, "Clustering Short Time Series Gene Expression Data," *Bioinformatics*, vol. 21, pp. i159-i168, 2005.

[13] T.J. Hestilow and Y. Huang, "Clustering of Gene Expression Data Based on Shape Similarity," *EURASIP J. Bioinformatics Systems Biology*, article 12, 2009.

[14] A. Ben-Dor and Z. Yakhini, "Clustering Gene Expression Patterns," *Proc. Third Ann. Int'l Conf. Computational Molecular Biology*, pp. 33-42, 1999.

[15] E.P. Xing and R.M. Karp, "Cliff: Clustering of High-Dimensional Microarray Data via Iterative Feature Filtering Using Normalized Cuts," *Bioinformatics*, vol. 17, no. suppl 1, pp. S306-S315, 2001.

[16] R. Sharan, A. Maron-Katz, and R. Shamir, "Click and Expander: A System for Clustering and Visualizing Gene Expression Data," *Bioinformatics*, vol. 19, no. 14, pp. 1787-1799, 2003.

[17] X. Wu et al., "Top 10 Algorithms in Data Mining," *Knowledge Information Systems*, vol. 14, no. 1, pp. 1-37, 2008.

[18] J.B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proc. Fifth Berkeley Symp. Math. Statistics and Probability*, pp. 281-297, 1967.

[19] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[20] S. Datta and S. Datta, "Comparisons and Validation of Statistical Clustering Techniques for Microarray Gene Expression Data," *Bioinformatics*, vol. 19, no. 4, pp. 459-466, 2003.

[21] I.G. Costa, F.A.T.d. Carvalho, and M.C.P. de Souto, "Comparative Analysis of Clustering Methods for Gene Expression Time Course Data," *Genetics Molecular Biology*, vol. 27, pp. 623-631, 2004.

[22] A. Thalamuthu, I. Mukhopadhyay, X. Zheng, and G.C. Tseng, "Evaluation and Comparison of Gene Clustering Methods in Microarray Analysis," *Bioinformatics*, vol. 22, pp. 2405-2412, 2006.

[23] M. Pirooznia, J. Yang, M.Q. Yang, and Y. Deng, "A Comparative Study of Different Machine Learning Methods on Microarray Gene Expression Data," *BMC Genomics*, vol. 9, no. Suppl 1, article S13, 2008.

[24] M.C.P. de Souto, I.G. Costa, D. de Araujo, T. Ludermir, and A. Schliep, "Clustering Cancer Gene Expression Data: A Comparative Study," *BMC Bioinformatics*, vol. 9, no. 1, article 497, 2008.

[25] E. Freyhult, M. Landfors, J. Onskog, T. Hvidsten, and P. Ryden, "Challenges in Microarray Class Discovery: A Comprehensive Examination of Normalization, Gene Selection and Clustering," *BMC Bioinformatics*, vol. 11, no. 1, article 503, 2010.

[26] R. Xu and D. Wunsch, *Clustering*. John Wiley & Sons, 2009.

[27] P.N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, first ed. Addison Wesley, 2005.

[28] K.W. Boyack et al., "Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches," *PLoS ONE*, vol. 6, no. 3, article e18029, 2011.

[29] C. Pesquita, D. Faria, H. Bastos, A.E.N. Ferreira, A.O. Falcão, and F.M. Couto, "Metrics for GO Based Protein Semantic Similarity: A Systematic Evaluation," *BMC Bioinformatics*, vol. 9, article S4, 2008.

[30] P.A. Jaskowiak, R.J.G.B. Campello, T.F. Covões, and E.R. Hruschka, "A Comparative Study on the Use of Correlation Coefficients for Redundant Feature Elimination," *Proc. 11th Brazilian Symp. Neural Networks (SBRN)*, pp. 13-18, 2010.

[31] R. Balasubramaniyan, E. Hullermeier, N. Weskamp, and J. Kamper, "Clustering of Gene Expression Data Using a Local Shape-Based Similarity Measure," *Bioinformatics*, vol. 21, pp. 1069-1077, 2005.

[32] Y.S. Son and J. Baek, "A Modified Correlation Coefficient Based Similarity Measure for Clustering Time-Course Gene Expression Data," *Pattern Recognition Letters*, vol. 29, pp. 232-242, 2008.

[33] A. Schliep, I.G. Costa, C. Steinhoff, and A. Schonhuth, "Analyzing Gene Expression Time-Courses," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 2, no. 3, pp. 179-193, July-Sept. 2005.

[34] C. Möller-Levet, F. Klawonn, K.-H. Cho, H. Yin, and O. Wolkenhauer, "Clustering of Unevenly Sampled Gene Expression Time-Series Data," *Fuzzy Sets and Systems,* vol. 152, pp. 49-66, 2005.

[35] K. Pearson, "Contributions to the Mathematical Theory of Evolution. III. Regression, Heredity, and Panmixia," *Proc. Royal Soc. London,* vol. 59, pp. 69-71, 1895.

[36] A. Brazma and J. Vilo, "Gene Expression Data Analysis," *FEBS Letters,* vol. 480, no. 1, pp. 17-24, 2000.

[37] R. Loganantharaj, S. Cheepala, and J. Clifford, "Metric for Measuring the Effectiveness of Clustering of DNA Microarray Expression," *BMC Bioinformatics,* vol. 7, no. Suppl 2, article S5, 2006.

[38] I. Priness, O. Maimon, and I. Ben-Gal, "Evaluation of Gene-Expression Clustering via Mutual Information Distance Measure," *BMC Bioinformatics,* vol. 8, no. 1, article 111, 2007.

[39] R. Steuer, J. Kurths, C.O. Daub, J. Weise, and J. Selbig, "The Mutual Information: Detecting and Evaluating Dependencies between Variables," *Bioinformatics,* vol. 18, pp. 231-240, 2002.

[40] I.G. Costa, F.A.T. Carvalho, and M.C.P. de Souto, "Comparative Study on Proximity Indices for Cluster Analysis of Gene Expression Time Series," *J. Intelligent Fuzzy Systems,* vol. 13, pp. 133-142, 2002.

[41] R. Gentleman, B. Ding, S. Dudoit, and J. Ibrahim, "Distance Measures in DNA Microarray Data Analysis," *Bioinformatics and Computational Biology Solutions Using R and Bioconductor,* pp. 189-208, Springer, 2005.

[42] R. Giancarlo, G. Lo Bosco, and L. Pinello, "Distance Functions, Clustering Algorithms and Microarray Data Analysis," *Proc. Fourth Int'l Conf. Learning and Intelligent Optimization,* C. Blum and R. Battiti, eds., pp. 125-138, 2010.

[43] R. Giancarlo, G. Bosco, L. Pinello, and F. Utro, "The Three Steps of Clustering in the Post-Genomic Era: A Synopsis," *Proc. Seventh Int'l Conf. Computational Intelligence Methods for Bioinformatics and Biostatistics,* pp. 13-30, 2011.

[44] M. Ashburner et al., "Gene Ontology: Tool for the Unification of Biology," *Nature Genetics,* vol. 25, no. 1, pp. 25-29, May 2000.

[45] P.A. Jaskowiak, R.J.G.B. Campello, and I.G. Costa, "Evaluating Correlation Coefficients for Clustering Gene Expression Profiles of Cancer," *Proc. Seventh Brazilian Symp. Bioinformatics (BSB '12),* pp. 120-131, 2012.

[46] L. Goodman and W. Kruskal, "Measures of Association for Cross-Classifications," *J. Am. Statistical Assoc.,* vol. 49, pp. 732-764, 1954.

[47] R.J.G.B. Campello and E.R. Hruschka, "On Comparing Two Sequences of Numbers and Its Applications to Clustering Analysis," *Information Sciences,* vol. 179, no. 8, pp. 1025-1039, 2009.

[48] M.G. Kendall, *Rank Correlation Methods,* fourth ed. Griffin, 1970.

[49] C. Spearman, "The Proof and Measurement of Association between Two Things," *Am. J. Psychology,* vol. 100, no. 3/4, pp. 441-471, 1904.

[50] D.J. Hand and R.J. Till, "A Simple Generalisation of the Area under the ROC Curve for Multiple Class Classification Problems," *Machine Learning,* vol. 45, pp. 171-186, 2001.

[51] C. Pesquita, D. Faria, A.O. Falcão, P. Lord, and F.M. Couto, "Semantic Similarity in Biomedical Ontologies." *PLoS Computational Biology,* vol. 5, no. 7, article e1000443, July 2009.

[52] K. Faceli, A.C.P.L.F. de Carvalho, and W.A. Silva Jr., "Evaluation of Gene Selection Metrics for Tumor Cell Classification," *Genetics and Molecular Biology,* vol. 27, pp. 651-657, 2004.

[53] P. Tamayo et al., "Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation," *Proc. Nat'l Academy Sciences USA,* vol. 96, no. 6, pp. 2907-2912, 1999.

[54] A.P. Gasch et al., "Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes," *Molecular Biology of the Cell,* vol. 11, no. 12, pp. 4241-4257, 2000.

[55] Z.S. Qin, "Clustering Microarray Gene Expression Data Using Weighted Chinese Restaurant Process," *Bioinformatics,* vol. 22, no. 16, pp. 1988-1997, 2006.

[56] P.T. Spellman et al., "Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast *Saccharomyces Cerevisiae* by Microarray Hybridization," *Molecular Biology Cell,* vol. 9, pp. 3273-3297, 1998.

[57] S. Chu et al., "The Transcriptional Program of Sporulation in Budding Yeast," *Science,* vol. 282, no. 5389, pp. 699-705, 1998.

[58] Y.H. Yang, S. Dudoit, P. Luu, and T.P. Speed, "Normalization for cDNA Microarray Data," *Microarrays: Optical Technologies and Informatics.* SPIE, pp. 141-152, 2001.

[59] P. Resnik, "Semantic Similarity in a Taxonomy: An Information-Based Measure and Its Application to Problems of Ambiguity in Natural Language," *J. Artificial Intelligence Research,* vol. 11, pp. 95-130, 1999.

[60] J.L. Sevilla et al., "Correlation between Gene Expression and GO Semantic Similarity," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 2, no. 4, pp. 330-338, Oct.-Dec. 2005.

[61] H. Frohlich, N. Speer, A. Poustka, and T. BeiSZbarth, "Gosim - An R-package for Computation of Information Theoretic Go Similarities between Terms and Gene Products," *BMC Bioinformatics,* vol. 8, no. 1, article 166, 2007.

[62] N. Bolshakova, F. Azuaje, and P. Cunningham, "A Knowledge-Driven Approach to Cluster Validity Assessment," *Bioinformatics,* vol. 21, no. 10, pp. 2546-2547, 2005.

[63] N. Bolshakova, A. Zamolotskikh, and P. Cunningham, "Comparison of the Data-Based and Gene Ontology-Based Approaches to Cluster Validation Methods for Gene Microarrays," *Proc. IEEE 19th Int'l Symp. Computer-Based Medical Systems (CBMS),* pp. 539-543, 2006.

[64] G. Milligan, "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms," *Psychometrika,* vol. 45, pp. 325-342, 1980.

[65] J. Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *J. Machine Learning Research,* vol. 7, pp. 1-30, 2006.

**Pablo A. Jaskowiak** received the BSc degree in informatics from the Western Paraná State University and the MSc degree from the University of São Paulo in 2008 and 2011, respectively. Since 2011, he has been working toward the PhD degree in computer science in the Institute of Mathematics and Computer Science at the University of São Paulo. His research interests are in the areas of machine learning and bioinformatics.

**Ricardo J.G.B. Campello** received the BSc degree in electronics engineering from the State University of São Paulo, Brazil, in 1994, and the MSc and PhD degrees in electrical engineering from the School of Electrical and Computer Engineering at the State University of Campinas, Brazil, in 1997 and 2002, respectively. Since 2007, he has been with the Department of Computer Sciences at the University of São Paulo, where he is currently an associate professor. His current research interests fall primarily into the areas of machine learning, data mining, and soft computing.

**Ivan G. Costa** received the bachelor's and MSc degrees in computer science from the Federal University of Pernambuco (UFPE), in 2001 and 2003, respectively, and the PhD degree from the Free University of the Berlin/ Max Planck Institute for Molecular Genetics in 2008. He is currently an assistant professor at UFPE. His research interests include gene expression analysis, personalized medicine, and statistical learning methods.