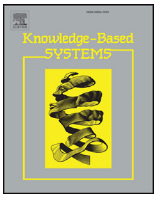
Knowledge-Based Systems 155 (2018) 83–87

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Original Software Publication

JBiclustGE: Java API with unified biclustering algorithms for gene expression data analysis

Orlando Rocha[a,b,*], Rui Mendes[b]

[a] Center of Biological Engineering, University of Minho, Portugal
[b] Department of Informatics, University of Minho, Portugal

ARTICLE INFO

Article history:
Received 19 December 2017
Revised 20 April 2018
Accepted 21 April 2018
Available online 25 April 2018

Keywords:
Biclustering
Gene expression
Data mining

ABSTRACT

Over the last years, comparative studies of biclustering algorithms have been described in the literature, showing that a variety of programming languages were used in their development. Because of this fact, many researchers have difficulty using some of these methods, since it is necessary to setup an environment for running a given algorithm or to have some programming skills in order to compile it. We present a new Java API for biclustering analysis in the context of gene expression data, allowing the use of 21 biclustering algorithms, in a single application. It is freely available at https://jbiclustge.github.io as an open source framework.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Since the advent of high-throughput measurements, such as DNA microarrays and next-generation sequencing technologies, researchers have tried to extract useful information from the cells and to understand how genes and their products interact under certain conditions and environmental limitations.

Clustering techniques have been used to identify and group a set of items (genes or conditions) into clusters, relying on distance and similarity functions to group items into subsets that exhibit a similar profile. These techniques have proven to be helpful in the discovery of biologically important groups of genes or samples, thus unveiling functional aspects of the biological processes [1,2]. However, these traditional clustering approaches restrict genes or conditions to belong to only one cluster. These assumptions can be misleading because genes may have a common behavior under a certain number of environmental conditions, and may have an independent behavior in other conditions. In view of these drawbacks, Cheng&Church [3] introduced the biclustering technique, based on work presented by Hartigan [4], in order to cluster subsets of genes and conditions simultaneously from gene expression data. Since then, many other biclustering methods have been proposed. The emergence of a large number of methods is due to the fact that the biclustering problem is NP-hard [5], thus leading to the development of different concepts and strategies of the algorithms to improve the insights of biologically relevant occurrences.

Several comparative studies have been carried out in the last years, describing the strengths and weakness of such algorithms in literature [5–10].

In this paper, we present a new application programming interface (API) developed in the Java language, that comprises 21 biclustering methods that can be used in the context of gene expression data analysis. The motivation that led to the implementation of this tool is the fact that in most of the comparative studies, the authors had to assemble their own research environment using the various methods implemented in several programming languages. For researchers with low programming skills, it can be difficult to use a large part of the existing algorithms, with the aggravating circumstance that each method uses a specific format for the results, making the comparison between them a hard task.

2. Problems and background

Several biclustering methods have been published in the last years. However, only a few biclustering frameworks, integrating more than one method, have been proposed. The biclustering analysis toolbox (BicAT) [11] is one of these frameworks that integrates six biclustering methods, and two standard clustering procedures. Furthermore, this tool has additional features that helps users perform biclustering analysis, such as: preprocessing, visualizing results and postprocessing functions. The biclust [12] R package is another well-known framework that integrates a small number of biclustering methods. This tool also provides some additional

[*] Corresponding author at: Centre of Biological Engineering, University of Minho, Portugal.

E-mail addresses: orocha@deb.uminho.pt (O. Rocha), rcm@di.uminho.pt (R. Mendes).

https://doi.org/10.1016/j.knosys.2018.04.024
0950-7051/© 2018 Elsevier B.V. All rights reserved.

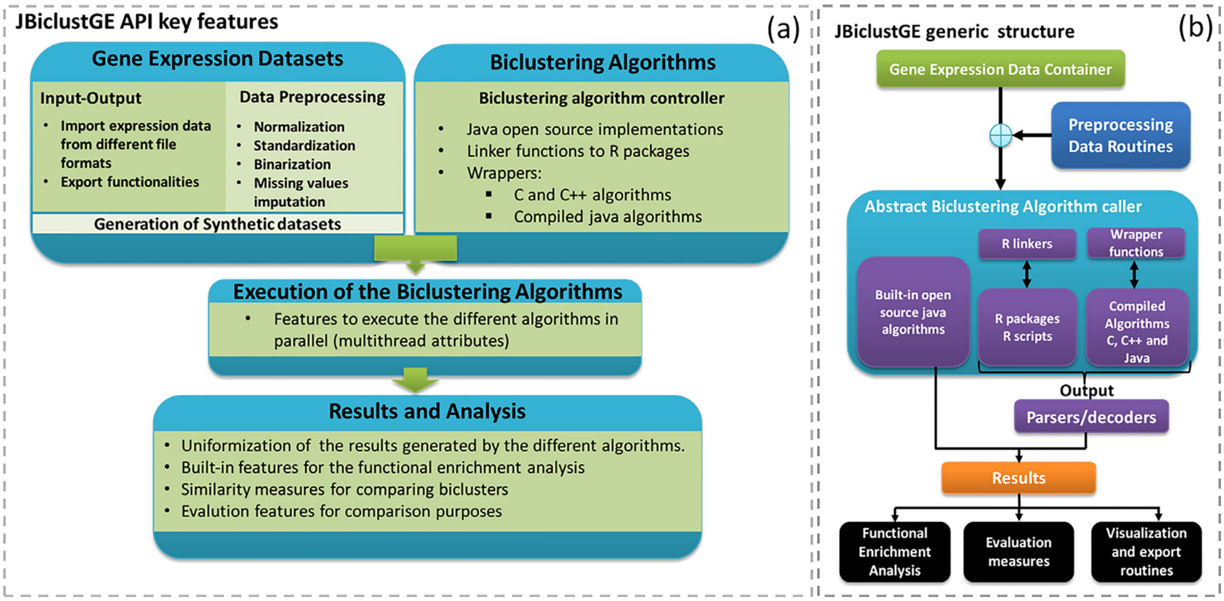

JBiclustGE API key features
(a)
Gene Expression Datasets
Input-Output
Import expression data from different file formats
Export functionalities
Data Preprocessing
Normalization
Standardization
Binarization
Missing values imputation
Generation of Synthetic datasets
Biclustering Algorithms
Biclustering algorithm controller
Java open source implementations
Linker functions to R packages
Wrappers:
C and C++ algorithms
Compiled java algorithms
Execution of the Biclustering Algorithms
Features to execute the different algorithms in parallel (multithread attributes)
Results and Analysis
Uniformization of the results generated by the different algorithms.
Built-in features for the functional enrichment analysis
Similarity measures for comparing biclusters
Evalution features for comparison purposes
JBiclustGE generic structure
(b)
Gene Expression Data Container
Preprocessing Data Routines
Abstract Biclustering Algorithm caller
R linkers
Wrapper functions
Built-in open source java algorithms
R packages R scripts
Compiled Algorithms C, C++ and Java
Output
Parsers/decoders
Results
Functional Enrichment Analysis
Evaluation measures
Visualization and export routines

Fig. 1. (a) Main implemented key features. (b) Generic structure of the JBiclustGE.

functionalities for the analysis and visualization of the results. Recently, Eren et al. [8] developed the BiBench platform to perform a comparative study of various methods. This tool was developed in the Python language and comprises twelve biclustering algorithms. In addition, evaluation functionalities for performing biclustering comparative analysis were implemented. In [13] Gupta et al. presented the MATLAB toolbox for biclustering analysis (MTBA), that includes twelve biclustering methods and multiple functionalities for data handling, preprocessing and visualization. Notwithstanding, most of these frameworks force users to have programming skills to use them. With the exception of BicAT none of the other frameworks have a user friendly graphical interface (GUI). However, BicAT includes only six biclustering algorithms. Moreover, MTBA is developed under MATLAB, an expensive proprietary platform. Table 2 presents a comparison between the features provided by JBiclustGE and the frameworks referred above. Thus, with the development of JBiclustGE, we intend to give and simplify the access to a large number of biclustering algorithms to the researchers without programming expertise and also provide several features for the analysis of the results of such algorithms, by providing a free and open source software.

3. JBiclustGE framework

The developed Java API comprises 21 methods for the biclustering analysis of gene expression data. These algorithms were chosen based on the following requirements: open source implementations or publicly available software implementations. In the development of this tool we used several open source Java libraries to attain all features presented in Fig. 1(a). Fig. 1(b) shows the generic structure of JBiclustGE concerning the interconnection of the implemented features, in order to perform the biclustering analysis of gene expression data.

3.1. Gene expression data features

The *Statistical Machine Intelligence and Learning Engine* (Smile) Java library was integrated in this API to afford the functionalities for data import, missing value imputation and data preprocessing. However, additional specific preprocessing methods had to be implemented for supporting the execution of some of the integrated biclustering algorithms. These implementations were performed according to the description of the mathematical routines provided by the authors in the literature. Expression data can be loaded from different file formats, such as: gene cluster text file format, ExpRESsion file format, Stanford cDNA file format, Text file format for expression dataset and attribute-relation file format. Subsequently, data can be normalized, scaled or binarized using one of the methods implemented in this tool. Moreover, synthetic datasets can also be generated, with the possibility of creating bicluster patterns (e.g. constant, constant with adjustments, plaid, shift, scaled and shift-scaled patterns) by following the procedures presented in [8,14,15], and also with added overlap and noise attributes.

3.2. Integration of the biclustering algorithms

This API integrates biclustering algorithms developed in the Java, R, C and C++ programming languages. Thus, different routines had to be implemented in order to support the input requirements, execution and the output results of these algorithms, using the strategies presented below. Table 1, shows the biclustering algorithms that were integrated in this tool, their availability and which type of routines were implemented to call/execute each of the presented algorithms.

3.2.1. Call routines of binary executables

Several routines had to be implemented in order to execute the binaries of the algorithms developed in C and C++, and the algorithms assembled in compiled JAR packages (Java Archive). These executables are called by passing all the required parameters via command-line interface, and the obtained results are converted using parsing routines developed specifically for each integrated algorithm.

3.2.2. Call routines through Rsession

Some of the biclustering algorithms are available as R packages. Thus, several routines had to be created in order to execute these algorithms in the R environment and to gather the results pro-

Download English Version:

https://daneshyari.com/en/article/6861331

Download Persian Version:

https://daneshyari.com/article/6861331

Daneshyari.com