ISSN No:-2456-2165

A Review of Visualization Methods and Tools for the Biclustering

Melih Sözdinler
Department of Computer EngineeringBogaziçi University
Bebek, Istanbul 34342 Turkey

Abstract:- Understanding the underlying information for interpreted data from raw data is currently a key challenge to Bioinformatics. This is essentially related to the current flood of raw data, aggregate information, and evolving knowledge arising from the study of gene expressions. For the analysis and understanding of that underlying information, biclustering takes the attention of authors from several different disciplines. As a result, hundreds of different variation of the biclustering problem has been introduced as a solution to biclustering problem. Although biclustering is well studied, there is still some gap for combined and interpreted visualizations methods. With this review paper, we are aimed to discuss the visualization methods of biclustering by comparing current the state of the art with each other. Additionally, we concentrate on the tools that provide visualizations. Through this paper, we first present the visualization methods. Then, we evaluate each proposed tools with the state of the art visualization options. Finally, we discuss future directions for biclustering visualization.

I. INTRODUCTION

The initial study of biclustering is first introduced by Hartigan [1], and presented the topic as the specialized version of *clustering* problem. In essence, clustering refers to the process of organizing a set of input vectors into clusters based on specified similarity with respect to some predefined distance measure [2]. In some cases having the relaxed types of clusters that can group input vectors both horizontally and vertically or in other words, *co-clustering* both features and samples are more appreciated. This special instance of clustering, named as *biclustering*. In general, clustering can have a resulting intuition that can be valuable in terms of a global perspective. The local perspectives and correlations are somehow disregarded by clustering algorithms unless the Principal Component Analysis or other dimensionality reduction methodologies are applied. Using biclustering, both global and local perspectives can be obtained by arranging the size of biclusters.

Biclustering continues to get attention from the research community. This is because there is still the existence of new opportunities to extend the problem and takes the attention from sub-disciplines of computer science, mathematics and statistics. As a result, this leads to several application areas such as data mining, pattern recognition, micro-array analysis, drug activity analysis, and motif detection [3], [4], [5],

Melih Sözdinler is supported by The Scientific and Technological Research Council of Turkey (TUBITAK) [BIDEB-2211].

[6], [7], [8], [9], [10], [11], [2], [12], [13], [14], [15], [16], [17], [18], [19], [20]. Particularly, biclustering problem turns into the optimization problem in several research papers. These papers mainly concentrate on solving the specified optimization problem. In most cases, the problem is NP- Hard and heuristic solutions are needed. Many heuristic opportunities make the topic up-to-date. Refer to these survey papers [21], [22], [23] for more details.

Since the topic is still attractive, the visualization and interpretation of the results of biclustering need more attention. In that way, supporting visualization results with P-Value [24], [25], [26], [2], H-value or Mean Squared Residue Score(MSRS)[11], Hv-value [12], Average Correlation Value(ACV) [27], Pearson Correlation Coefficient [28], are common for biclustering research papers. In most cases, biological interpretation is supported with visualization, and underlying information is demonstrated.

In this review paper, we give the visualization methods and state of the art of existing tools. Then, we compare the pros and cons of these tools. In the literature, we see the biclustering algorithm explosion in 21 years starting with Cheng and Church [11]. Vice versa, visualizations options for biclustering are limited and open to new methods. During the review, we also suggest that existing specialized visualization methods are more beneficial as supporting evidence of the quality of biclusters. For this purpose, we try to discuss novel visualization methods throughout the paper. We expect that this review provides new directions for the forthcoming studies.

Recently, several tools for biclustering that have embedded supports for the visualization options. The main visualization options are Heatmap Representation [29], [30], [31], [32] and Parallel Coordinate Plots [29], [30], [31], [32], [33]. Rather than these, there are also specialized visualization models. These are Force Directed Layout Model [32], Bubble Map Model [32], Mountain Map Model [34], Enrichment Tree Visualization [30], Integrated Visualization of Biclusters Verifying with Protein-Protein Interactions [35], Advanced Heatmap Representation [31] and Modified Parallel Coordi- nate Visualizations [30], [33], [32]. BiCluster Viewers [36] is also be applied to highlight detected biclusters generated from the original data set by using heatmaps and parallel coordinate plots as visualization methods. BiCFlows [37] also provided a novel approach to the visualization of bipartite graphs

ISSN No:-2456-2165

where this also well fits into bicluster visualization. The tool allows for multi-scale exploration through the hierarchical aggregation of nodes and edges using biclustering in the linked lists. Different than all these visualization methods, Forestogram [38] specifically considers Hierarchical Biclusters and they provide a 3D method inspired by dendrograms to visualize biclusters.

The Furby [39] is another interactive visualization tech- nique for analyzing biclustering results. They have twofold contribution claims. The first one is a high-level view of the overall results. And, that shows bicluster shared rows and columns with visualization. The second contribution is to provide heatmaps and bar charts and enable analysts to interactively set the thresholds that transform the fuzzy (soft)clustering into hard clusters.

In this review paper, we first discuss the visualization methods. Then, we overview proposed tools and their pros and cons. Finally, we talk about future directions, suggestionsand discussions.

II. VISUALIZATION METHODS

Through the last section, we review the recent literature on visualization methods where there are several different approaches. We divide these into simple and hybrid methods. Simple visualization methods are mainly introduced before biclustering in the literature. On the other hand, hybrid methods are the extensions of existing models, and in most cases, they are completely new methods. Furthermore, some visualization approaches not only provide bicluster-specific visualizations but also give some other insights into bicluster relations. The resulting visualization snapshot as one main layout and supporting sub-layouts have some interactivity. On the other hand, the remaining approaches consider an all-in-one approach which means that for each bicluster, they give the corresponding visualizations and there are no such high-level visuals.

In this section, we review the visualization methods intwo sub-section. First, we present single methods, next we demonstrate hybrid methods.

A. Simple Methods

Heatmap Representation and Parallel Coordinate Visualiza- tion are assumed as traditional hence they introduced before the biclustering essentially for data visualization. In general, they are useful because you can obtain basic knowledge aboutthe biclusters by looking at the resulting pictures of these approaches.

Heatmaps show us the selected parts of the dimensions and values corresponding to these selected dimensions of the dataset, also referred to as biclusters. The main contribution of heatmaps is their contribution to infer information about the bicluster structure. By looking at the resulting heatmap, we see the structure of a bicluster and we can predict the type of a bicluster such as *all constant*, *constant row*, *constant column*, *coherent*, etc. Particularly, the structure of the bicluster in heatmap depends on the optimization criteria and scoring metric of the algorithm. Therefore, heatmaps are one of the basic validation of the results. A Colouring scheme is vital for heatmaps. Mainly, variations of colors (green-red) are used and each color represents the interval of values. Additionally, choosing an appropriate interval is necessary as well in order to evenly distribute the colors. Heatmaps are included in several tools [30], [29], [40], [18], [32], [31], [35] and each tools have some different modifications.

Parallel Coordinates(PC) plots are another useful and common visualization method of biclusters. PC plots mainly introduce visuals for one dimension of data with respect to other dimensions. From the perspective of gene expression data, the PC plots represent the genes under the subset of conditions. Using PC plots, one can easily detect the bicluster structure and it is easier to see *coherent* biclusters than heatmaps. In PC plots, if the overall picture has simultaneous increasing and decreasing plots, this means that the corresponding bicluster has a correlation in itself. Furthermore, PC plots need to have an appropriate coloring scheme and scaling of plots. Some approaches also add fuzzy effects on each plot to simplify the complicated drawings. Also, scaling is vital to show peak points, intervals of values. Parallel Coordinates are supported in several tools, makes it fundamental for biclustering tools [30], [40], [29], [33], [16], [32], [35]. Note that, some of these tools provide PC plots as *Gene Expression Profiles* [29], [30] which is the non-scaled version of PC plots. We will not distinguish these similar approaches through the review.

Indeed, these two simple methods for visualizations are supplementary and most of the cases, these are not enough. As an example, in case of relatively sized biclusters, simple visualizations methods can not represent visual intuition perfectly.

B. Hybrid Methods

The common feature of hybrid methods are targetting thebiological relevance and providing the general picture of the results while proposing a new idea or extending the existing methods. We review these methodologies in this subsection. The first visualization approach is based on Force Directed Layout [32]. The authors' claim is to unravel trends and to highlight relevant genes and conditions using both visualapproaches and complementing biological and statistical analysis. Hence, end-users may have a chance to explore the results quickly and interactively. The visualization technique obtains its root from force-directed layouts that is represented as flexible overlapped groups of genes and conditions. Themodel is integrated with Heatmaps and Parallel CoordinatePlots and its advantage is the availability of the extension of its visualization methodology with biological relevance using transcriptional modules. Moreover, their proposed model shows several biclusters at once that also combines the overallview with the traditional approaches. Additionally, in theirmodel, if the nodes are connected, a spring force keeps thenodes closer. Also, there is an expansion force that pushesevery pair of nodes whether connected or not [41]. Theirclaim is that nodes in the same biclusters are closer and the remaining nodes belonging to different

ISSN No:-2456-2165

biclusters are separated. To represent the bicluster as a graph, they form a complete graph for each bicluster. For overall visibility, they do not show all the edges and nodes. Instead, they show the hull of each bicluster with some transparency. The transparency of this hull is useful to show overlapping biclusters and to increase visibility.

The second approach is a tree-based visualization method that gives biological relevance to biclusters. The methodmakes a connection between functional categories of an organism in Gene Ontology [42] and the corresponding biclusters. It is proposed by [30]. In their method, they form a Gene Ontology(GO) category tree for each bicluster. The tree is in hierarchical layout and specification increases at each lower level of the layout. Next, they calculate Bonferroni corrected P-value overall category nodes of the constructed tree. They color the GO categories on the behalf of lower p-values. In addition to that, they have also different colors to show GO main categories; *cellular component*, *molecular function* and *biological process*. The intensity of each color changes according to the calculated P-value. The main contribution of this approach is to integrate the biclustering concept with the fair scoring of P-values. Rather than this approach, Func Associate [24] and several other tools have to support to the calculation of P-values. With [30] approach, we observe these results with a well-readable picture rather than text and we are able to see the hierarchy of GO categories inside the visual. The origin of the approach depends on the hierarchical layout using a well-known graph drawing tool named graphviz [43]. Since graphviz provides nice visualizations, the picture of the layout is well readable and useful to detect enriched categories. Additionally, a similar central graph, scores of each bicluster are calculated and nodes or biclusters in the the main graph have a size proportional to these scoring functions. They use three scoring functions; *H-value* [11], *Hv-value* [12] and Enrichment Ratio similar to *P-value*. Finally, edges in the central graph also show either the common genes or interacting edges between two biclusters. The advantage of IntegratedViz is its integration of PPI networks and biclusters, to show the biological relevance of each set of genes of biclusters in global and local views. With this idea, correlated biclusters tend to have more interactions in their corresponding PPI networks.

Another approach in [31], extends the traditional heatmap visualization. The proposed idea shows all biclusters on a special type of heatmap. The methodology extends the heatmap layout by including multiple labels of genes and conditions in the resulting heatmap. So, they are able to show each bicluster while maintaining the minimum number of repetitions of labels. They propose a novel algorithm, based on *PQ-Tree*. The algorithm is based on finding an ordering such that the binary formation of 1's at each bicluster is consecutive. This is called *Consecutive One's Property (COP)*. In the beginning, discretization of bicluster data is needed in the form of 0's and 1's. Then, they set-up *PQ-Trees* from each discretized bicluster M and next, its rows are stored in list L. Using *REDUCE* operation, they perform hierarchical clustering to maintain COP property. Next, using the *MERGE* operation, they form resulting *PQ-Trees* as a new list, L' by looking at the similarity score of column lists C_T and $C_T *$ where T and T^* are separate *PQ-Trees*. The similarity score is as follows.

Approach is proposed as in the histogram format [40]. Each histogram shows both the significant transcription factors and GO functional enrichment analysis of bicluster genes. Users can easily detect the most common transcription factor or most common category according to the calculated P-value using the histogram.

$$\sigma(T, T^*) = \frac{C_T \cap C_T *}{C_T \cup C_T *} \quad (1)$$

At next, An integrated model for visualizing biclusters from gene expression data and PPI networks (Integrated Viz) the tool is introduced [35]. They proposed an approach that integrates Protein-Protein Interaction (PPI) networks and biclusters. Their method has one central graph that represents each bicluster as a node and each node in the main graph has a peripheral graph that corresponds to the sub-network formed by genes of the bicluster. The biological relevance is maintained by using these sub-networks of each bicluster and the edges between nodes of central graph. The Layout is based on *Weighted Hierarchical Layout*. Authors propose this layout as an extension to *Unweighted Hierarchical Layouts* as a new graph drawing approach. Introducing weights into graphs give more options to enrich the layout to demonstrate extra weight information of edges. Furthermore, comprising the peripheral graphs are done such that genes of the corresponding biclusters are extracted from the PPI network and eventually, using weighted hierarchical layout algorithm the final layout is obtained. Peripheral graph edges between genes show the reliability of the interactions. At the $\sigma(T, T^*)$ is a function to decide merge operation and each σ function results for all pairs are calculated at the beginning and sorted. Then, they perform *REDUCE* operation between the pairs with the highest σ. If it fails, *MERGE* operation doesnot occur. In *REDUCE* operation, basically, they are checking that the restrictions defined by *PQ-Tree*, T, holds for T^*. If it holds, *MERGE* operation occurs. T and T^* are deleted from L. Merged tree T_m is added to list by upgrading σ values. Finally, when all σ values are processed, they give the final layout as it appears on the set of columns of *PQ-Trees*. This method provides a combinatorial algorithm that holds for the minimum number of repetitions of labels where they are part of the original data in the resulting heatmap. Their method works better in overlapping biclusters and makes it available to show several biclusters in one heatmap. In several overlapping bicluster cases, their problem, as they mentioned, is the number of biclusters. To avoid this problem, they develop a web-based interface that allows execution and navigation through the web. There is also another specialized methodology on heatmap visualization in [44]. Although they mainly discuss the clustering point of view, their proposedtoolbox should be applicable to biclustering. In that approach, they extend the view of heatmaps into the third dimension using dendrograms and it is more desirable in such a case to see all biclusters in one visualization.

ISSN No:-2456-2165

In addition to all these mentioned approaches, in [33], they propose an extended parallel coordinate visualization. Their approach is to give parallel coordinate plots of a bicluster by simultaneously drawing with the real data plots and the bicluster plots of the same conditions. They are colored with different colors and bicluster gene plots are more visible to emphasize. According to this claim, the global view of the parallel plots is not hidden. This provides a better understanding of gene plots over a subset of conditions. Furthermore, in [30], it is not a brand new method, but they have one screen that includes all plots. They provide all bicluster plots together to speed up the plots extraction process and this gives us to look at several results to investigate both local and global patterns of biclusters. Finally, in [32] they perform some improvements to obtain good scaling and coloring.

Bubble Map is also another simple approach used in the literature, but in biclustering, it has a special meaning to show the projections of Mountain Maps [34]. This way, we can represent the bicluster as 2D bubbles with different sizes.Due to the method suggestion, we accept it as innovative ones. In the method, each bubble can have meaning such as correlated patterns inside the biclusters and higher values in sub-spaces. 3*D* version of bubble map is mountain map. Mountain maps are proposed for cluster visualization [34]. In [32], they propose bubble map representation as a projection of 3D mountains in 2D as bubble maps.

In the conclusion of this section, visualization methods provide us with both local and global perspectives of biclusters. In recent trends, integrating biclusters with biological analysis and data becomes more popular and next-generation visualization tools should include this integration. We now discuss the existing Analysis Methods.

III. SURVEY OF EXISTING TOOLS

In this section, we aim to review features of the existing tools and the visualization and the analysis methods are outlined here.

In Table I, we give an overview of popular visualization tools for biclustering. We provide brief overview with this table. Other than visualization tools, there are also tools proposed for the self-execution of some algorithms [18], [45], [46], [47] where those can be the subject of another review paper. Rest of the subsections, state of the art tools in Table I is going to be presented with their pros and cons.

A. Expander

Expander (EXpression Analyzer and DisplayER) [40], [4],

[48] is the eldest tool in our review and still get updates. Expander is a complete solution and supports heatmap and PC visualizations, bicluster analysis, and execution of algorithms. The tool supports the execution of their own biclustering algorithm SAMBA and clustering algorithm CLICK. Expander also provides the visualization of heatmaps, the resulting bitmaps can be saved easily, and biological evaluation in terms of functional categories in GO using corrected P-value is available. As designed, the histogram of biclusters gives the specific encountered categories. Hence this property supports the biclustering results. The tool has other options such as Principal Component Analysis, viewing Box Plots and recently added "Network Based Grouping". Additionally, Expander allows saving sessions which is also useful to continue at the point saved. The overall snapshot of Bicluster related visualizations and other features described in Figure 1. Expander is implemented with JAVA and free access is available 1.

B. Biclustering Analysis Toolbox(BicAT)

BicAT [29] is one of the early integrated tools for both ex- ecution, analyzing, and visualization of biclusters. It supports heatmap and PC visualizations. The main contribution of the tool is to provide a framework for the execution of well-cited algorithms CC [11], OPSM [3], ISA [7] and XMOTIF [5] and Bimax [9]. On the other hand, the tool has no biological supported visualizations and narrow analysis support. In general, it is one of the earliest tools for bicluster visualization and it is famous because of its variety of supported algorithms and simple GUI. BicAT is coded in JAVA and it is free and downloadable[2]. However, supported algorithms inside the tool are binary only. The Heatmap and PC example from BicAT provided in Figure 2.

C. BiVoc

BiVoc [31] specializes in the heatmap representation by extending the representation as multiple biclusters over the input matrix. The algorithm based on *PQ-Trees* is explained in the previous section. It is innovative since there is no such work to show overlapping biclusters as one heatmap for biclusters. The support of tools is limited due to specialization. They are supporting a defined input format for submitting biclusters and navigation via a web-based interface. Their main concern is to visualize the overlapping biclusters. The methodology of BiVoc does not concentrate on having a unique label. One label in the resulting heatmap could be represented several times but they claim that this repetition is minimized with the guaranteed algorithm. Therefore, for less number of biclusters, their method gives fine results in terms of the total number of rows and columns on the layout. Vice versa, despite the minimization, when the number of biclusters is high, their method gives several rows and columns that may disturb the overall view. Their solution to this problem is providing a web-based interface to follow and track the results. The program is coded in C++ and free access is available[3].

D. BiVisu

BiVisu [33] is proposed with the algorithm called PM. Their contribution is mainly on the PC plot drawing. Their approach draws parallel coordinate plots by giving plots of bicluster genes within plots of all genes inside the data. The color plots correspond to genes of bicluster with a different color. This enables the user to see the global view of the parallel gene plots for the corresponding bicluster with respect to a set of conditions in the bicluster. The tool also

ISSN No:-2456-2165

providesa heatmap view. Further analysis of biclusters are available in its GUI such as *H-value* and *Average Correlation Value*. Although they propose an extension to PC plots, the view hassome problems of scaling as shown in [49]. The Heatmap and PC example from BiVisu provided in Figure 3. Their program is implemented in MATLAB[4].

1Expander: EXpression Analyzer and DisplayER
2BicAT: Biclustering Analysis Toolbox
3BiVoc: Automatic layout and visualization of biclusters

	Heatmap	PC	Specialized Approach	Algorithm Support	Biological Evaluation	Bicluster Analysis
BicAT 2006	Yes	Yes	No	OPSM, Bimax, CC, XMOTIF, ISA	No	Yes
BiVoc 2006	Yes	No	New Heatmap to visualize all biclusters	Import interface for existing results	No	No
BiVisu 2007	Yes	Yes	Modified PC plots	Own Algorithm	No	Yes
Bicoverlapper 2008	Yes	Yes	Force Directed Layouts	Import interface for existing results	Yes	Yes
BiGGEsTS 2009	Yes	Yes	Functional Category Tree View	ECCC	Yes	Yes
IntegratedViz 2010	No	No	Integrated Visualization with PPI networks	Import interface for existing results, REAL, Bimax, CC	Yes	Yes
Robinviz 2011	Yes	Yes	A Graph Based, Reliability Measurement Criteria added upon IntegratedVis and Gene Ontology categories integrated to the results of biclustering algorithms	REAL, Bimax, CC	Yes	Yes
Furby 2014	Yes	Yes	Force Directed approach provides main graph consists of biclusters and showing the actual data that forms the individual clusters together with the information which rows and columns they share	FABIA	Yes	Yes
Forestogram 2017	Yes	No	A visualization tool helps practitioners to understand how biclusters evolve	Hierarchical biclustering	Yes	Partially
Expander 2003 and 2019	Yes	Network based grouping	No	SAMBA	Yes	Yes

Table 1 : Overview of existing tools

A. Bicoverlapper

Bic Overlapper [32], [49] is one of the sophisticated tools that mainly concentrates on the visualization techniques that exist before. They also propose a brand new method. In this new method, they use the force-directed layout of a graph of corresponding biclusters. The detail of the method is given in the previous section. Since it provides visualization for several biclusters, their method differs from other visualization methods except BiVoc and Integrated Viz do. Their main contribution is handling several overlapping and non over lapping biclusters with given biological relevance with respect to Transcriptional Regulatory Networks(TRN). They also support their main layout with heatmaps and PC plots as evidence of their integrated visualization, and they propose the 2D Bubble Map method by applying from 3D version named Mountain Map visualization. Moreover, they do not have implemented algorithms inside their tool, however, they provide the import interface for the results of biclustering algorithms. Their main concern is the execution time of force- directed layouts. These layouts are simple and easy to apply when the graph has a countable number of nodes. In the case of biclustering results with higher dimensions, meaning that many nodes, force-directed graph may slow the execution and meaning of the main layout may be intervened. Their tool is available online5 and they provided some example biological analysis in the paper [49]. The overall visualization snapshot of BicOverlapper provided in Figure 4. Sample result for their unique method for showing Overlapping biclusters is given at Figure 4-a.

[4]BiVisu: Bicluster Visualization

ISSN No:-2456-2165

B. BiGGEsTS

BiGGEsTS (Biclustering Gene Expression Time Series) [30]is another state of the art tool that supports both several visualizations and analysis methods for biclusters. The tool also maintains the execution of their algorithm [10]. Its GUI is similar to BicAT and thus user-friendly. They provide embedded visualization methods which are heatmap, PC, multiple PCs, and enrichment tree visualization based on the method described in the last section. BiGGEsTS maintainsan integrated environment. You can obtain both heatmaps

[5]BicOverlapper: Bicluster Overlapper

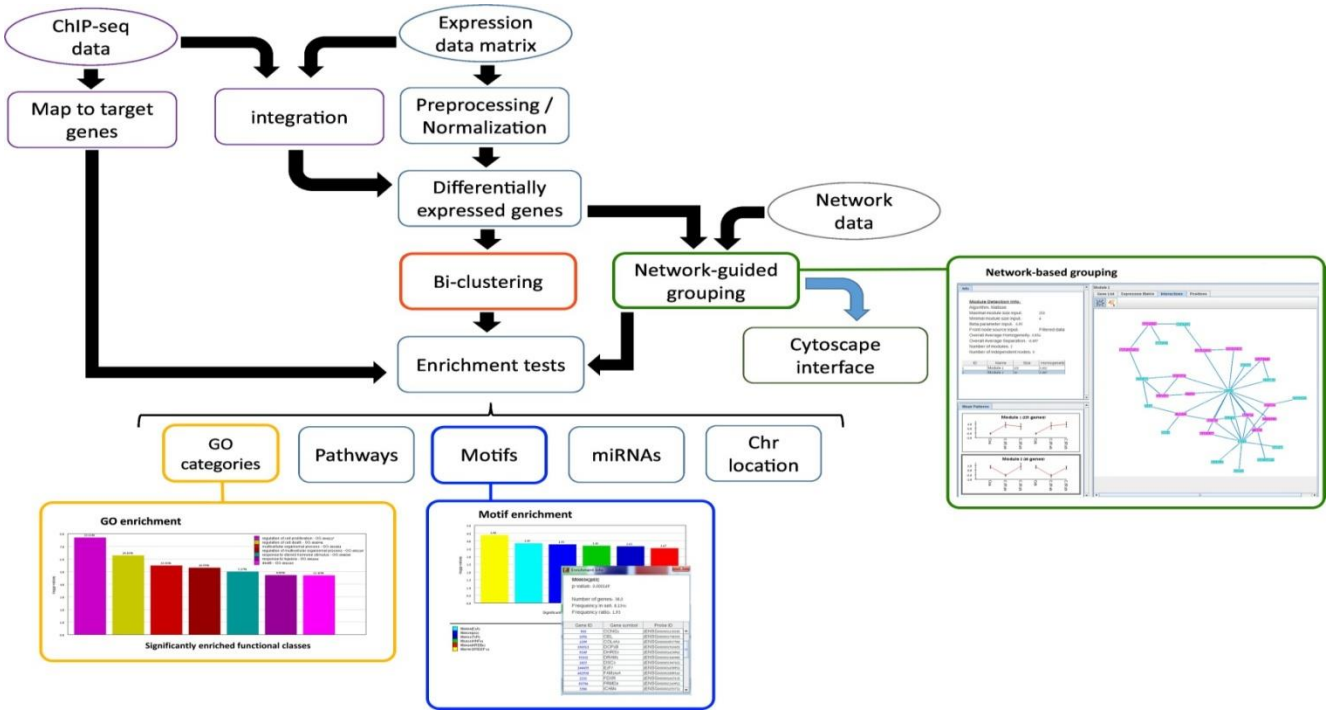


Fig. 1: Expander Features in [48]

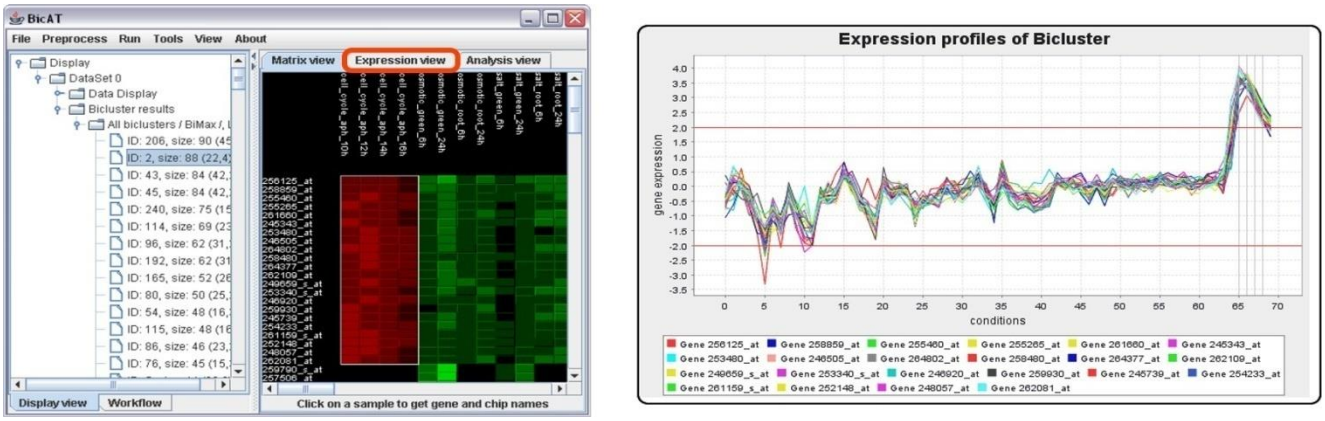


Fig. 2: Bicat's Heatmap and Parallel Coordinate Support in [29]

ISSN No:-2456-2165

and PCs. Also, their innovative enrichment tree visualization methodology gives the biological significance of biclusters. Furthermore, in their multiple PC plots, they also give the expression patterns that simplify the congested plots. These simplified plots show the general trend of PC plots and the pattern of bicluster as well. Additionally, BiGGEsTS use the Graphviz tool to support their enrichment tree visualization. The whole procedure except the production of graphs is executed using Graphviz's dot program. BiGGEsTS supports sessions to save all the work that is done during the session and also allows execution of their own algorithm eCCC. The Heatmap and PC examples from BiVisu provided in Figure 5. Their enrichment tree view using Gene Ontology [42] main categories is also given inside the figure. This approachcan help research teams to compare the biclusters with real clustered domains. Finally, BiGGEsTS is implemented usingJAVA and license is under GPL[6].

[6]BiGGEsTS: Biclustering Gene Expression Time Series

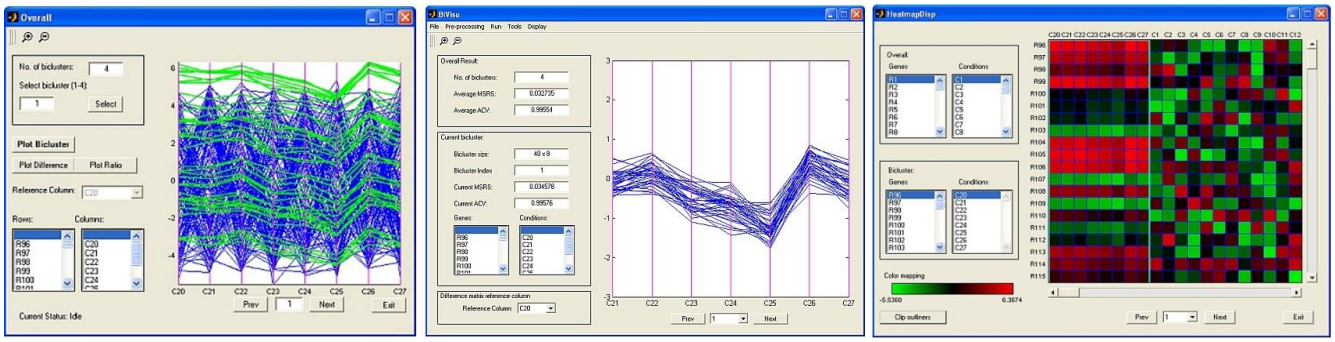
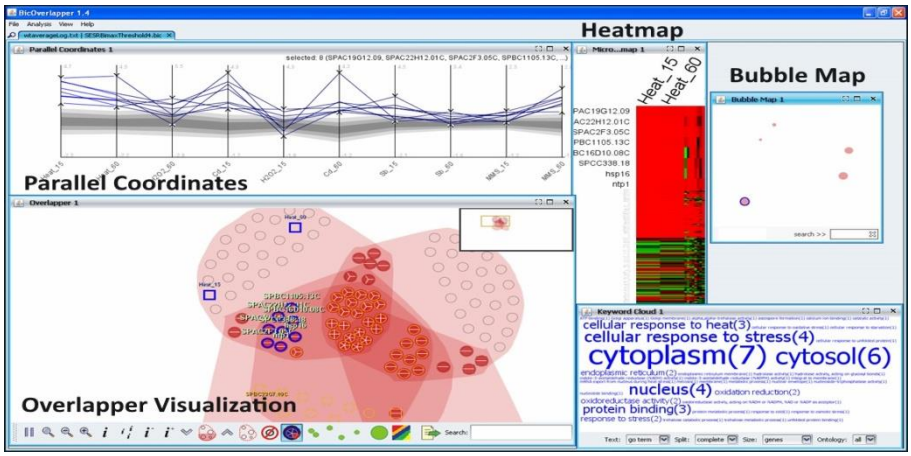
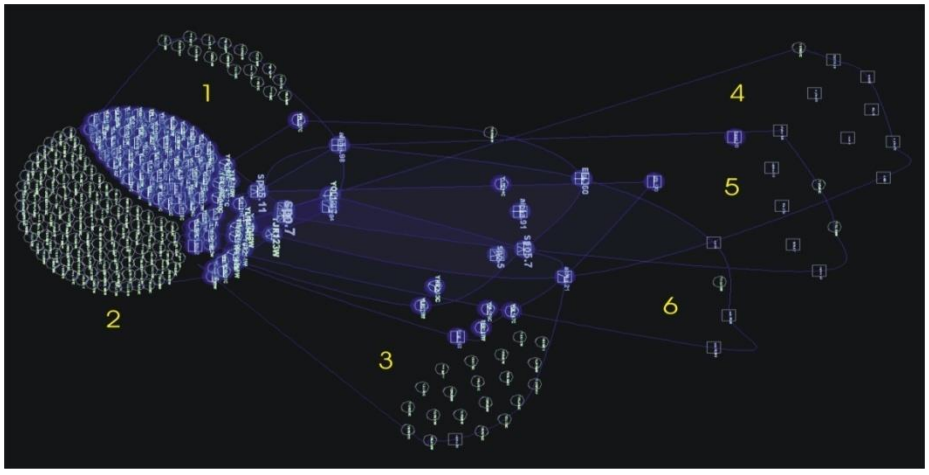


Fig. 3: BiVisu's Heatmap and Parallel Coordinate Support in [33]



(a)



(b)

Fig. 4: (a) Bicoverlapper's Overall UI in; (b) Overlapper visualization for Biclustering Result of OPSM's coherent evolution biclusters, finding only about a dozen in Saccharomyces Cerevisiae example[49]

ISSN No:-2456-2165

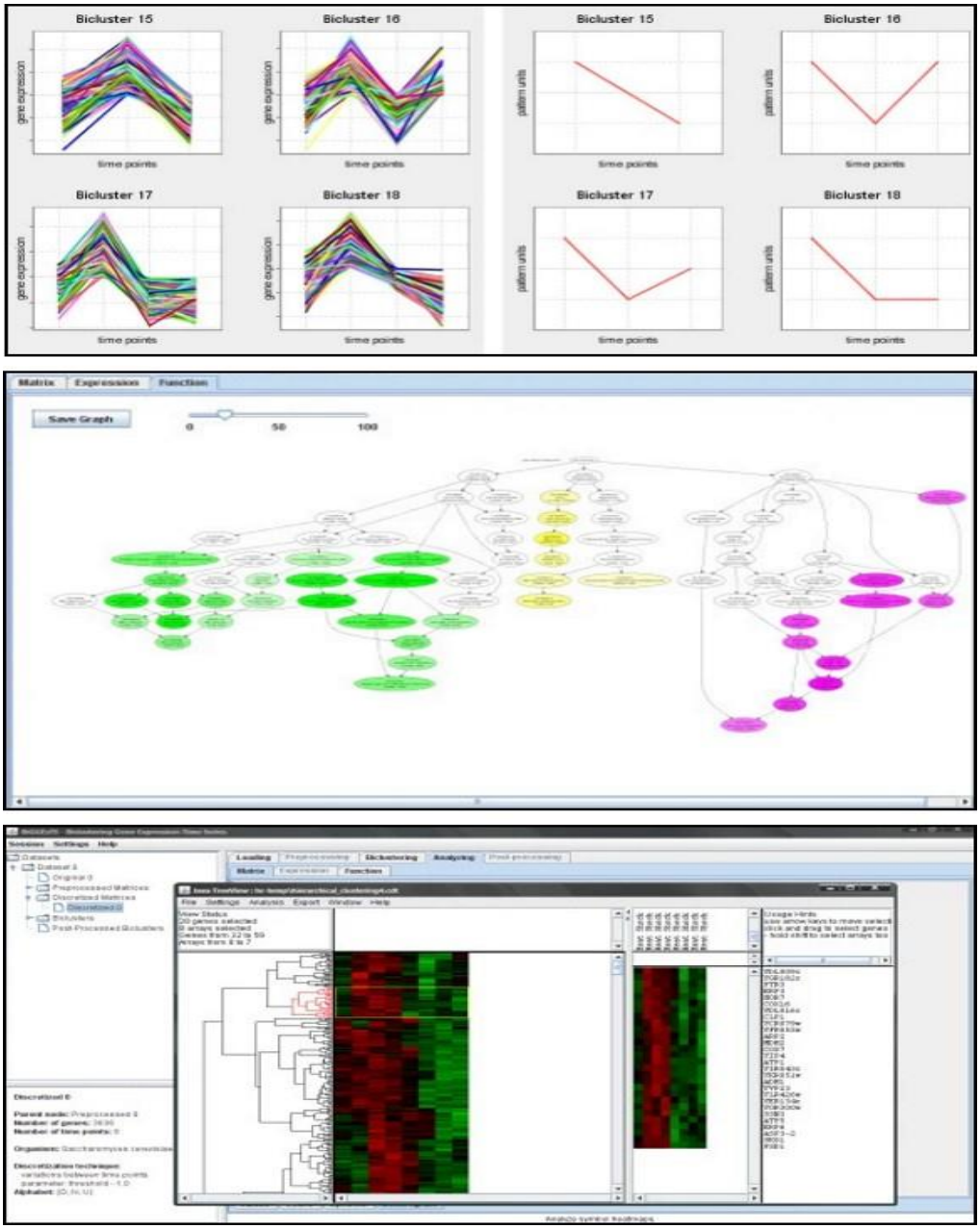


Fig. 5: BiGGEsTS's Visualization Methods[30]; Parallel Coordinates Support, Gene Ontology Tree Function Visualization,Heatmap's with Dendograms

C. IntegratedViz and Robinviz

Integrated Viz is an integrated visualization tool and proposes an innovative approach of evaluation and validation of biclustering results using biological data. Integrated Viz also supports Heatmap and PC plots visualizations. Particularly, Integrated Viz concentrates on both global and local visualization of biclusters. Global view shows each bicluster as a node of weighted hierarchical layout and peripheral graphs corresponding to these graphs are accessible via clicking. The details of the methodology are described in the previous section. Due to the nature of the proposed methodology, IntegratedViz also provides some pieces of evidence for analysis such as scoring and enrichment value of biclusters. Furthermore, their visualization method is also supported with visual clues such as coloring of categories at peripheral graphs which shows the main category of genes among pre- determined ones, edge thickness, and node sizes. The main graph shows the H-values of each bicluster by arranging the size of nodes. They also allow importing the results of algorithms and execution of given algorithms in Table I.

ISSN No:-2456-2165

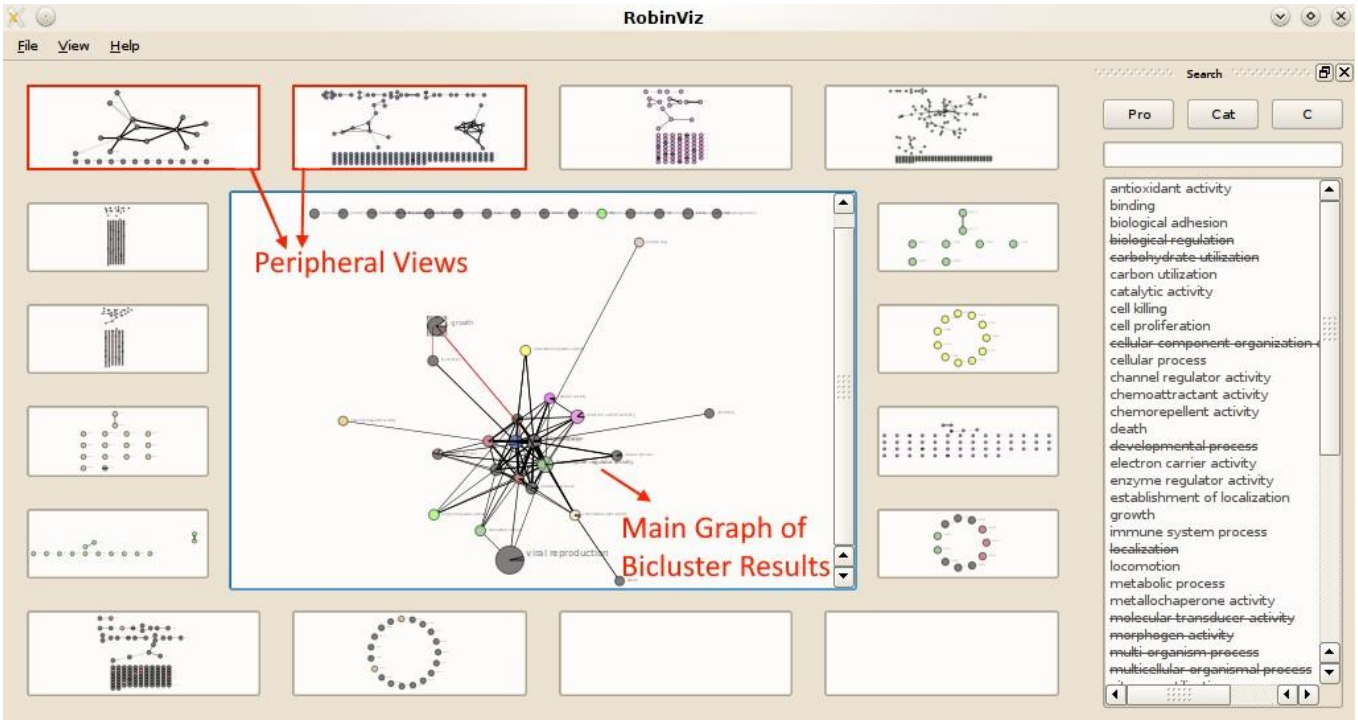


Fig. 6: Robinviz's Visualization Methods[50]; Integrated function of Peripheral Views and Main Bicluster Result Graph

The one disadvantage of the tool is again the problem of several biclusters as it happens in BicOverlapper and BiVoc. To prevent visualization disturbance, they added some scaling and hiding methods. Integrated Viz is written in C++ and freely available[7]. They also extended this tool and renamed as Robinviz (Reliability Oriented Bioinformatics Network Visualization) [50][8]. The Figure 6 shows peripheral views and main graph of bicluster results. And, the proposed visualization shows bicluster relations as graph with weightednodes and edges.

D. Furby

Furby (Fuzzy Force-Directed Bicluster Visualization)[39] yields two-fold contribution. First, they target to achieve a high-level graph that contains each biclusters as node and edges are defined accordingly similarities of biclusters. Second, for fuzzy bi-clustering results, their technique lets analysts set the thresholds interactively. In that way, such transformation from the fuzzy (soft) clustering into hard clusters can then be investigated using heatmaps or bar charts. Furby can import biclustering results applied to a multi-tissuedataset and show them into the visualization[9]. Similar toRobinviz, Furby also provides bicluster correlations within the main user interface. Inside the Figure 7 running example provided. In this example, the main visualization graph is also supported with Heatmaps, and that approach supports in- terfering with the bicluster structure with visualization. In addition to that, weighted edges provide additional insight into the relationship between two biclusters. The tool is also designed to limit some edges so that the resulting graph is more simple to visualize and layout.

[7]IntegratedViz: Integrated Model for Visualizing Biclusters

[8]Robinviz: Reliability Oriented Bioinformatics Network Visualization

[9]Furby: Fuzzy Force-Directed Bicluster Visualization

E. Forestogram

Forestogram[38] iteratively collects rows or columns and uses them to construct a forest. This mainly relies on the collection of dendrograms with a common root. In Figure 8, we provide a sample execution scenario of their methods. Their visualization method also supports Hierarchical biclus- tering mainly. In Figure 8, their method is demonstrated with a simple example; from data matrix to the construction of forestogram. Within the thesis of the author, forestogram is also used for the biological dataset. The visualization feature of forestogram yields a comprehensive tool for analyzing the co-related features across the pregnancy trimesters. And this is used for the interpretation of the results.

IV. FUTURE DIRECTIONS AND CONCLUSION

Indeed, we give a review of existing visualization methods and tools. For the methods, we review *Simple* and *Hybrid* methods with a short explanation to each of them. Then, We also demonstrate the state of the art of this domain with proposed tools. We mainly review specialized approaches specific to visualization. And in Table I an overview of thesetools written.

As an observation, the number of research papers related to biclustering algorithms is increasing every year. However, the visualization methods are limited. As a consequence of this, there are so many bicluster finding methods but the dilemma is rather than common methods for visualization, further interpretation methods are limited.

ISSN No:-2456-2165

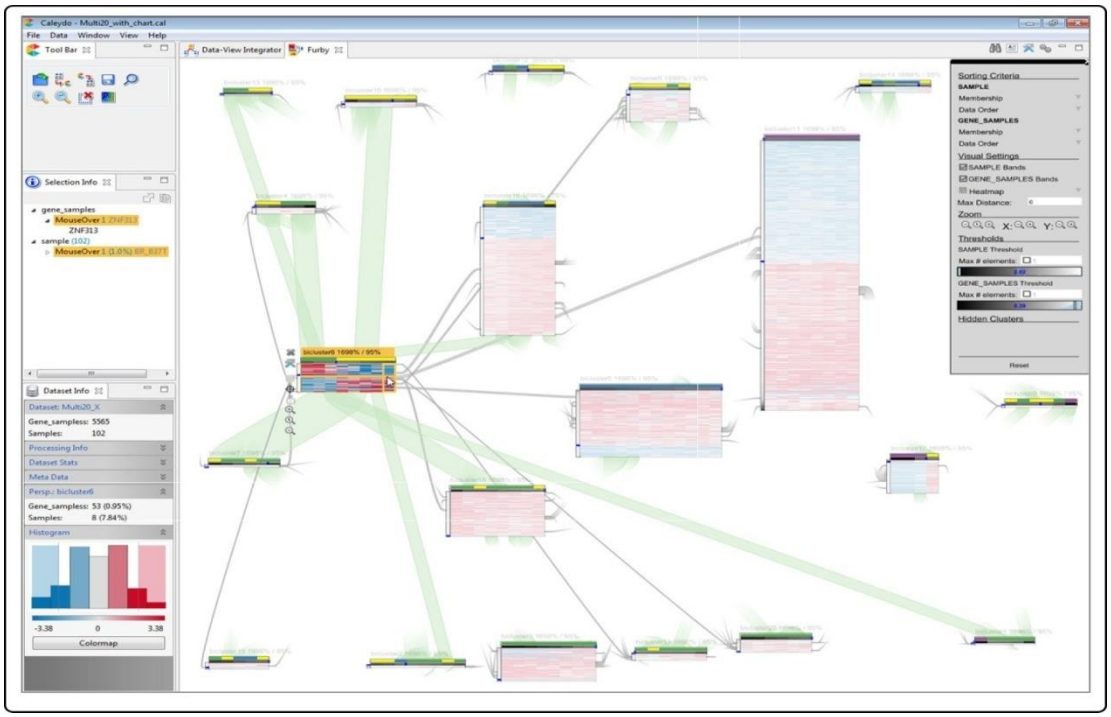


Fig. 7: Furby's Visualization Methods[39]; main force-directed visualization of the graph supplemented with Heatmaps andedge weighs

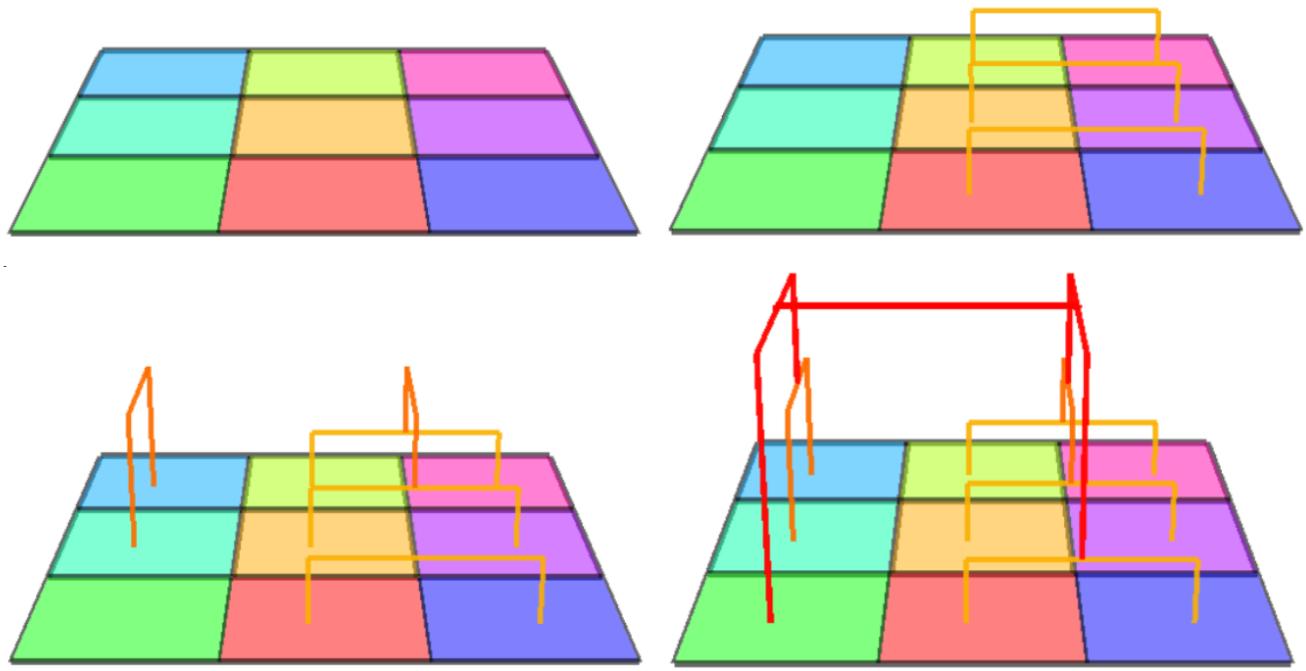


Fig. 8: Forestogram Algorithm Execution [38]; the initial data matrix, pair of columns and merge, then, a pair of rows andmerge, and the final completed forestogram

```
# Sample Result File
# Algorithm : CC
BiclusterNumber: 2
7 3
gene 1 gene 5 gene 11 gene 12 gene 14 gene 15 gene 16
cond 1 cond 3 cond 5
3 5
gene 2 gene 3 gene 4
cond 2 cond 4 cond 6 cond 7 cond 8
# Algorithm Parameters:
MaxRows: 10
MinRows: 2
MaxColumns: 10
MinRows: 2
```

Fig. 9: Suggested Bicluster Result Output

Additionally, the validation of the visualizations with otherbiological data such as Functional Categories, PPIs and TRNs surely be important in next-generation bicluster visualizations and analyses. Heatmaps and Parallel Coordinate Plots are helpful commonly. However, these methods are not self enough to demonstrate the quality using some biological relevant data. Since gene expression data are the main input of biclustering algorithms, the validation of the existing algorithms is surely done with the biological relevance of genes. In that case, visualization methods should be able to show some clues about these relevance to inform end-users. In [51], they have a claim that networks in biology can appear complex and difficult to decipher. They provide a concept in order to analyze these networks using frequently employed visualization and analysis patterns. This approach supports our claim for the bicluster visualization. The increasing amount of data in subtopics of bioinformatics may result in relation to the biclustering problem. As a result, the visualization approaches from these topics can be integrated to decipher these relations.

For the future direction, some standardization is also needed for the community. By now, we have some common standards for gene expression data and tools or biclusering algorithm implementation can easily parse this information. On theother hand, we observe that there is no common way for importing the results of biclusters. This is important since theproposed tools are based on the assumption that any input from biclustering algorithm can be used and rendered. We suggest that tools should not follow many different bicluster result formats. Indeed the new algorithm source codes can output stable format. This can be a new mark-up language format that is easier to parse or even, using very simple notation format is also possible as shown inside Figure 9.

This notation provides the name of the algorithm, number of biclusters, size of genes and conditions at each biclus- ter, and the labels of genes and conditions in a bicluster. After these suggestions for biclustering results, in the end, biclustering algorithm-related parameters can be listed for the information of end-user and memorization. Surely, this notation can be the topic of another paper for providingcommon unique file format for the biclusters.

Furthermore, providing open-source codes or self-executable programs for the proposed biclustering algorithms should be beneficial. Therefore, it is possible to look at the other proposed method without reimplementation. Especially in biclustering, it is more desirable to have open-source codes or self-executable programs to help the prospective authors for their comparisons. Additionally, existing tools can publish their code at GitHub. So that, the community can participate coding and forking may help the evolution of future tools.

In a conclusion, we believe the visualization of biclusters is still a hot topic. We especially focus on Bioinformatics data'sbiclustering and visualization. Through the review paper, wedemonstrate visualization methods and then we give future directions and suggestions. Understanding the biclustering results and providing a *Hybrid* visualization approaches by integration with other computer science topics will become more and more important in the future.

V. ACKNOWLEDGEMENTS

Melih Sözdinler is paid by The Scientific and Technologi- cal Research Council of Turkey(TUBITAK)[BIDEB-2211]. Additional thanks to my PhD advisor; Can Özturan for the support and patience.

Conflict of Interest: None declared.

REFERENCES

[1.] J. A. Hartigan, "Direct clustering of a data matrix," *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 123–129, 1972.

[2.] Erten and M. Sözdinler, "Biclustering expression data based on ex- panding localized substructures," in *Bioinformatics and Computational Biology*, S. Rajasekaran, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 224–235.

[3.] Ben-Dor, B. Chor, R. Karp, and Z. Yakhini, "Discovering local structure in gene expression data: the order-preserving submatrix problem," in *RECOMB '02: Proceedings of the sixth annual international conference on Computational biology*. New York, NY, USA: ACM, 2002, pp. 49–57. [Online]. Available: http://dx.doi.org/10.1145/565196.565203

[4.] Tanay, R. Sharan, and R. Shamir, "Discovering statistically significant biclusters in gene expression data." *Bioinformatics*, vol. 18 Suppl 1, 2002. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/12169541

[5.] T. M. Murali and S. Kasif, "Extracting conserved gene expression motifs from gene expression data," in *Pacific Symposium on Biocomputing*, 2003, pp. 77–88. [Online]. Available: http://helix-web.stanford.edu/psb03/murali.pdf

[6.] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, "Spectral biclustering of microarray data: coclustering genes and conditions." *Journal Genome Res PMID 12671006*, vol. 13, pp. 703–16, 2003. [Online]. Available: http://bioinfo.mbb.yale.edu/genome/expression

[7.] S. Bergmann, J. Ihmels, and N. Barkai, "Iterative signature algorithm for the analysis of large-scale gene expression data." *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 67, no. 3 Pt 1, March 2003. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/12689096

[8.] Abdullah and A. Hussain, "A new biclustering technique based on crossing minimization," *Neurocomputing*, vol. 69, no. 16-18, pp. 1882–1896, 2006. [Online]. Available: http://dx.doi.org/10.1016/j. neucom.2006.02.018

[9.] Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Buhlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, "A systematic comparison and evaluation of biclustering methods for gene expression data," April 2006. [Online]. Available: http://bioinformatics.oxfordjournals.org/cgi/content/short/22/9/1122?rss=1

[10.] S. Madeira and A. Oliveira, "A polynomial time biclustering algorithm for finding approximate expression patterns in gene expression time series," *Algorithms for Molecular Biology*, vol. 4, no. 1, p. 8, 2009. [Online]. Available: http://www.almob.org/content/4/1/8

[11.] Y. Cheng and G. M. Church, "Biclustering of expression data." in *ISMB*, P. E. Bourne, M. Gribskov, R. B. Altman, N. Jensen, D. A. Hope, T. Lengauer, J. C. Mitchell, E. D. Scheeff, C. Smith, S. Strande, and H. Weissig, Eds. AAAI, 2000, pp. 93–103. [Online]. Available: http://dblp.uni-trier.de/db/conf/ismb/ismb2000.html

[12.] K. Bryan and P. Cunningham, "Extending bicluster analysis to annotate unclassified orfs and predict novel functional modules using expression data." *BMC genomics*, vol. 9 Suppl 2, 2008. [Online].

[13.] Available: http://dx.doi.org/10.1186/1471-2164-9-S2-S20

[14.] J. Liu, Z. Li, X. Hu, and Y. Chen, "Biclustering of microarray data with mospo based on crowding distance." *BMC bioinformatics*, vol. 10 Suppl 4, no. Suppl 4, pp. S9+, 2009. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-10-S4-S9

[15.] Gyenesei, U. Wagner, S. Barkow-Oesterreicher, E. Stolte, and R. Schlapbach, "Mining co-regulated gene profiles for the detection of functional associations in gene expression data," *Bioinformatics*, vol. 23, no. 15, pp. 1927–1935, August 2007. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btm276

[16.] J. Gu and J. S. Liu, "Bayesian biclustering of gene expression data." *BMC genomics*, vol. 9 Suppl 1, no. Suppl 1, pp. S4+, 2008. [Online].

[17.] Available: http://dx.doi.org/10.1186/1471-2164-9-S1-S4

[18.] K.-O. Cheng, N.-F. Law, W.-C. Siu, and A. Liew, "Identification of coherent patterns in gene expression data using an efficient biclustering algorithm and parallel coordinate visualization," *BMC Bioinformatics*, vol. 9, no. 1, p. 210, 2008. [Online]. Available: http://www.biomedcentral.com/1471-2105/9/210

[19.] S. Dharan and A. Nair, "Biclustering of gene expression data using reactive greedy randomized adaptive search procedure," *BMC Bioinformatics*, vol. 10, no. Suppl 1, p. S27, 2009. [Online]. Available: http://www.biomedcentral.com/1471-2105/10/S1/S27

[20.] P. Carmona-Saez, R. Pascual-Marqui, F. Tirado, J. Carazo, and Pascual-Montano, "Biclustering of gene expression data by non- smooth non-negative matrix factorization," *BMC Bioinformatics*, vol. 7, no. 1, p. 78, 2006. [Online]. Available: http://www.biomedcentral.com/1471-2105/7/78

[21.] X. Gan, A. Liew, and H. Yan, "Discovering biclusters in gene expression data based on high-dimensional linear geometries," *BMC Bioinformatics*, vol. 9, no. 1, pp. 209+, April 2008. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-9-209

[22.] Li, Q. Ma, H. Tang, A. H. Paterson, and Y. Xu, "Qubic: a qualitative biclustering algorithm for analyses of gene expression data." *Nucleic acids research*, vol. 37, no. 15, pp. e101+, August 2009. [Online]. Available: http://dx.doi.org/10.1093/nar/gkp491

[23.] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: A survey." *IEEE/ACM Trans. on Comp. Biol. and Bioinformatics (TCBB)*, vol. 1, no. 1, pp. 24–45, 2004.

[24.] Tanay, R. Sharan, and R. Shamir, "Biclustering algorithms: A survey," *Handbook of Computational Molecular Biology*, 2004. [Online]. Available: http://en.wikipedia.org/wiki/Biclustering

[25.] S. Busygin, O. Prokopyev, and P. M. Pardalos, "Biclustering in data mining," *Comput. Oper. Res.*, vol. 35, no. 9, pp. 2964–2987, September 2008.

[Online]. Available: http://dx.doi.org/10.1016/j.cor.2007.01.005

[26.] F. Berriz, O. D. King, B. Bryant, C. Sander, and F. P. Roth, "Characterizing gene sets with funcassociate," *Bioinformatics*, vol. 19, no. 18, pp. 2502–2504, December 2003. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btg363

[27.] I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock, "Go::termfinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes." *Bioinformatics*, vol. 20, no. 18, pp. 3710–3715, December 2004. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/15297299

[28.] K. Bryan and P. Cunningham, "Bottom-up biclustering of expression data." *Proceedings of the 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB'06*, no. 4133177, pp. 232–239, 2006.

[29.] L. Teng and L.-W. Chan, "Biclustering gene expression profiles by alternately sorting with weighted correlated coefficient," in *Machine Learning for Signal Processing, 2006. Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on*, 2006, pp. 289–294. [Online]. Available: http://dx.doi.org/10.1109/MLSP.2006.275563

[30.] Bhattacharya and R. K. De, "Bi-correlation clustering algorithm for determining a set of co-regulated genes," *Bioinformatics*, vol. 25, no. 21, pp. 2795–2801, November 2009. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btp526

[31.] S. Barkow, S. Bleuler, A. Prelic, P. Zimmermann, and E. Zitzler, "Bicat: a biclustering analysis toolbox." *Bioinformatics (Oxford, England)*, vol. 22, no. 10, pp. 1282–1283, May 2006. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/16551664

[32.] J. Goncalves, S. Madeira, and A. Oliveira, "Biggests: integrated environment for biclustering analysis of time series gene expression data," *BMC Research Notes*, vol. 2, no. 1, p. 124, 2009. [Online]. Available: http://www.biomedcentral.com/1756-0500/2/124

[33.] Grothaus, A. Mufti, and T. M. Murali, "Automatic layout and visualization of biclusters," *Algorithms for Molecular Biology*, vol. 1, no. 1, pp. 15+, September 2006. [Online]. Available: http://dx.doi.org/10.1186/1748-7188-1-15

[34.] R. Santamaria, R. Theron, and L. Quintales, "BicOverlapper: A tool for bicluster visualization," *Bioinformatics*, vol. 24, no. 9, pp. 1212–1213, 2008. [Online]. Available: http://bioinformatics.oxfordjournals.org/cgi/content/abstract/24/9/1212

[35.] K.-O. Cheng, N.-F. Law, W.-C. Siu, and A. Liew, "Bivisu: software tool for bicluster detection and visualization." *Bioinformatics (Oxford, England)*, vol. 23, no. 17, pp. 2342–2344, September 2007. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btm338

[36.] M. Rasmussen and G. Karypis, "gcluto - an interactive clustering, visualization, and analysis system," Tech. Rep., 2004.

[37.] E. Aladağ, C. Erten, and M. Sözdinler, "An integrated model for visualizing biclusters from gene expression data and ppi networks," in *ISB '10: Proceedings of the 1th international symposium on Biocomputing*. Calcut,Kerala,India: ACM, 2010.

[38.] Heinrich, R. Seifert, M. Burch, and D. Weiskopf, "Bicluster viewer: A visualization tool for analyzing gene expression data," in *Advances in Visual Computing*, G. Bebis, R. Boyle, B. Parvin, D. Koracin, S. Wang,

[39.] Kyungnam, B. Benes, K. Moreland, C. Borst, S. DiVerdi, C. Yi-Jen, and J. Ming, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 641–652.

[40.] Steinbock, E. Groller, and M. Waldner, "Casual visual exploration of large bipartite graphs using hierarchical aggregation and filtering," in *2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA)*, 2018, pp. 1–10.

[41.] M. S. Ghaemi, V. Partovi Nia, and B. Agard, "Forestogram: A visualization framework for hierarchical biclustering," pp. 1–16, May 2017. [Online]. Available: https://www.gerad.ca/en/papers/G-2017-40

[42.] M. Streit, S. Gratzl, M. Gillhofer, A. Mayr, A. Mitterecker, and S. Hochreiter, "Furby: fuzzy force-directed bicluster visualization," *BMC Bioinformatics*, vol. 15, pp. S4 – S4, 2014.

[43.] R. Sharan, A. Maron-katz, and R. Shamir, "Click and expander: A sys- tem for clustering and visualizing gene expression data," *Bioinformatics*, vol. 19, pp. 1787–1799, 2003.

[44.] T. Kamada and S. Kawai, "An algorithm for drawing general undirected graphs," *Inf. Process. Lett.*, vol. 31, no. 1, pp. 7–15, 1989.

[45.] Reference Genome Group of the Gene Ontology Consortium, "The gene ontology's reference genome project: a unified framework for functional annotation across species." *PLoS computational biology*, vol. 5, no. 7, pp. e1 000 431+, July 2009. [Online]. Available: http://dx.doi.org/10.1371/journal.pcbi.1000431

[46.] Ellson, E. Gansner, E. Koutsofios, S. North, and G. Woodhull, "Graphviz and dynagraph – static and dynamic graph drawing tools," in *Graph Drawing Software*, M. Junger and P. Mutzel, Eds. Springer-Verlag, 2003, pp. 127–148.

[47.] S. Rajaram and Y. Oono, "Neatmap - non-clustering heat map alternatives in r," *BMC Bioinformatics*, vol. 11, no. 1, p. 45, 2010. [Online]. Available: http://www.biomedcentral.com/1471-2105/11/45

[48.] X. Liu and L. Wang, "Computing the maximum similarity bi-clusters of gene expression data," *Bioinformatics*, vol. 23, no. 1, pp. 50–56, 2007. [Online]. Available:

http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/1/50

[49.] A. Shabalin, V. J. Weigman, C. M. Perou, and A. B. Nobel, "Finding large average sub matrices in high dimensional data," *ANNALS OF APPLIED STATISTICS*, vol. 3, p. 985, 2009. [Online]. Available: doi:10.1214/09-AOAS239

[50.] S. Kaiser and F. Leisch, "A toolbox for bicluster analysis in r," 2008. [Online]. Available: http://epub.ub.uni-muenchen.de/3293/

[51.] T. A. Hait, A. Maron-Katz, D. Sagir, D. Amar, I. Ulitsky, C. Linhart, Tanay, R. Sharan, Y. Shiloh, R. Elkon, and R. Shamir, "The expander integrated platform for transcriptome analysis," *Journal of Molecular Biology*, vol. 431, no. 13, pp. 2398–2406, 2019, computation Resources for Molecular Biology. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022283619302803

[52.] R. Santamaria, R. Theron, and L. Quintales, "A visual analytics approach for understanding biclustering results from microarray data," *BMC Bioinformatics*, vol. 9, no. 1, pp. 247+, May 2008. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-9-247

[53.] A. E. Aladağ, C. Erten, and M. Sözdinler, "Reliability-Oriented bioinformatic networks visualization," *Bioinformatics*, vol. 27, no. 11, pp. 1583–1584, 04 2011. [Online]. Available: https://doi.org/10.1093/bioinformatics/btr178

[54.] Merico, D. Gfeller, and G. D. Bader, "How to visually interpret biological data using networks." *Nature biotechnology*, vol. 27, no. 10, pp. 921–924, October 2009. [Online]. Available: http://dx.doi.org/10.1038/nbt.1567