MILESTONE TRANSACTIONS ON
MEDICAL TECHNOMETRICS
VOL. 1, ISSUE. 1, 2023

RESEARCH ARTICLE

Bi-Cluster Based Analysis on Gene Ontology

Meenakshi Sundaram A[1] . Anooja Ali [1] . S S Patil [2] . Ajil A[1]

[1]School of Computer Science and Engineering, REVA University, Bangalore, Karnataka, India.
[2] Department of Agricultural Statistics University of Agriculture Sciences, Bengaluru, India

Received: 05 March 2023 / Revised: 12 April 2023 / Accepted: 13 June 2023
©Milestone Research Publications, Part of CLOCKSS archiving
DOI: 10.5281/zenodo.8073114

Abstract – Understanding biological activity requires the detection of crucial proteins. The identification of significant genes throughout the entire genome is advantageous for a number of reasons, including the categorization of critical genes for health and sickness, the rational creation of drugs, etc. Statistical methods have been suggested for predicting essential or requisite proteins/gene/GO terms, employed in protein networks. The computational approaches focusing on the topological characteristics or centrality approaches ignore the biologically relevant intrinsic features of essential proteins. Hence, considering the biological aspects like expression data, subcellular information, annotation data, and orthologous relationships can improve accuracy. So, in this research, bi-clustering algorithm is used to detect the essential Gene Ontology (GO) terms in molecular, cellular and biological processes by evaluating the protein associations and encoding the associations with ontology terms and pathways. The proposed method encodes each protein in terms of Mutual Information (MI) score, GO annotation and vector-based GO encoded matrix is generated and the essential proteins are extracted. The validation of the proposed method is verified using different statistical measures on the datasets.

Index Terms – Bicluster, Centrality, Gene Ontology, Mutual Information

I. INTRODUCTION

The field of bioinformatics has made significant progress as a prominent discipline by integrating computational and information science principles to extract knowledge from biological datasets[1]. A plethora of bioinformatics resources, databases, and software tools have been developed to facilitate the storage and processing of biological data. Proteins, essential molecules found in all living organisms, can undergo various conformational changes when interacting with other molecules [2]. However, existing proteomics tools currently lack the capability to perform comprehensive analysis of biological data, leading to insufficient prediction and function analysis. The presence of numerous irrelevant and noisy features in high-throughput proteomic results often obscures the identification of true indicators.

Bioinformatics facilitates the biological data to be stored as databases and bioinformatics tools enabled researchers to extract the biological data from the databases and analysis. The important Protein-Protein Interaction (PPI) repository includes Databases of Interacting Proteins (DIP) [3], Biological General Repository for Interaction Datasets (BioGRID) [4], and Search Tool for the Retrieval of Interacting Proteins/Genes (STRING) [5]. The remarkable advancement in bioinformatics is the result of integrating computational and predictive systems, which has motivated researchers to explore vast datasets and make significant advancements and discoveries through optimization.

Essential proteins can assist us understand how cells function and pinpoint the genes responsible for numerous illnesses in humans. Topological structures with centrality measurements can identify the crucial protein. It is crucial to take gene ontologies' biological characteristics into account[6]. Understanding the various viewpoints is made easier by capturing the ontology concepts. The suggested approach uses the GO annotation and MI score to encode the protein relationships. In this research, the authors implemented a biclustering method to detect the essential GO terms by encoding functional associations with pathway information [7]. Biclustering combined with annotation can efficiently identify key GO keywords. The remaining of the paper has been arranged as follows: section 2 briefs the literature survey and motivation for the analysis. Section 3 and 4 deals with methodology and results. Finally, the concluding remarks and references are specified.

II. LITERATURE REVIEW

The investigation of crucial proteins will lay the groundwork for maintaining life form, cellular functions, and the logical design of medicines, the recognition of disease genes, and the development of a cell in synthetic biology. Various statistical techniques have been proposed to predict essential or necessary proteins, genes, or GO terms, utilizing network topologies. There are diverse centralities measures available in the literature to uncover the essential proteins in a PPI network. These structure-based techniques concentrate on hubs and may be used to identify the influential node in any network [8]. The essentiality-lethality rule was developed as a result of the relationship between essentiality and lethality. Eigenvector centrality, degree, betweenness, proximity, and subgraph are some examples of the traditional centrality measurements. The research becomes more complicated and inconsistent when a benchmark centrality metric is not available on a certain network. Studies show, however, that having more information about orthologous connections can aid in efficiently identifying key proteins. Therefore, it is crucial to suggest a different method of identifying important proteins.

Centrality rule provides evidence the functional prominence more strongly with essentiality identification than topological centrality [9]. The conventional methods like, mutagenesis is often time consuming. So, taking into account biological data is the answer to the essentiality and topological controversy. By combining gene expression as the co-expressed neighbourhood with the topological characteristics of the PPI, Zhang proposed CoEWC [10], a metabolic route is created by a number of molecular connections, therefore grouping according to the pathway may be preferable. It is not necessarily true in bioinformatics that genes with similar expression patterns act in the same way. Proteins engage in numerous interactions and form complexes to carry out various biological functions. Consequently, assigning a specific biological function to an individual component becomes challenging.

Bioinformatics has emerged as a potent approach for controlling genes that possess multiple characteristics. Orzechowski demonstrated biclustering or co-clustering as an unsupervised method that allows simultaneous clustering of rows and columns along with the properties [11].

Bicluster are characterized as a group of genes and conditions under which these genes can participate in related biological processes in specific conditions. The main use of biclustering with annotated data is to identify local patterns that exist in the dataset. When a set of conditions is added, a gene branches with equal expression levels and is identified as a Bicluster [12]. Biclustering is an effective method for obtaining the results of gene annotation since it will group gene samples that are physiologically related. Cheng and Church developed the first biclustering program to discover patterns of local similarity in gene expression data [13]. This deterministic greedy method seeks to lower a submatrix's Mean Squared Residue (MSR) score. There exist a variety of biclustering methods and adopted numerous techniques and algorithmic principles that direct the search for concrete outcomes. Another approach, the Order Preserving Sub Matrix (OPSM), tries to build wide submatrices in which the levels of expression of all genes result in the same linear sample ordering [14]. In order to delve into the expression data, it is often necessary to perform analytical assessments and identify co-regulated proteins. Consequently, there is a need for more suitable and efficient computational methods to identify significant Gene Ontology (GO) terms. In the presented research, biclustering is employed to identify essential GO terms. An algorithm is implemented to detect essential proteins by encoding genes using the Kyoto Encyclopedia of Genes and Genomes (KEGG).

III. METHODOLOGY

Biclusters fall into three categories: coherent values, constant row values, and constant column values [15]. A coherent Bicluster has two variations: additive and multiplicative. Biclusters can be exclusively row-wise and column-wise, overlapping and non-overlapping. Biclustering methods follow metric and non-metric strategies. They perform search criteria- stochastic or iterative, greedy or heuristic accompanied by metric approaches. Figure 1 represents the proposed system architecture. The database known as STRING provides access to functional associations, with the most recent version being STRING 11.0. This database covers a wide range of 5090 organisms and supports genome enrichment analysis. The enrichment analysis is facilitated by utilizing GO and KEGG as classification models. Gene ontology models capture the dynamic nature of protein-protein interactions (PPIs) and reveal protein complexes. The protein associations represented in STRING correspond to biologically significant interfaces that contribute to the functional system in living organisms. As a benchmark dataset, STRING offers an analysis framework for functional and network enrichments.

QUalitative BIClustering algorithm, QUBIC is a qualitative or semi-qualitative deterministic algorithm to measure biclusters with discretized matrix [16]. The level of gene expression is quantified in QUBIC as an integer number in various circumstances. The incoming data matrix is transformed into a discrete numeric rank matrix during the data preparation stage. Yet, discretization could be excessively strict for datasets heavily impacted by noise, and the method might skip several rows and columns of a Bicluster. Since there is no noise in gene expression data, QUBIC is regarded as an appropriate biclustering approach. The biclust package is used to do statistical calculations by the package QUBIC-R, which is more efficient than the C implementation.

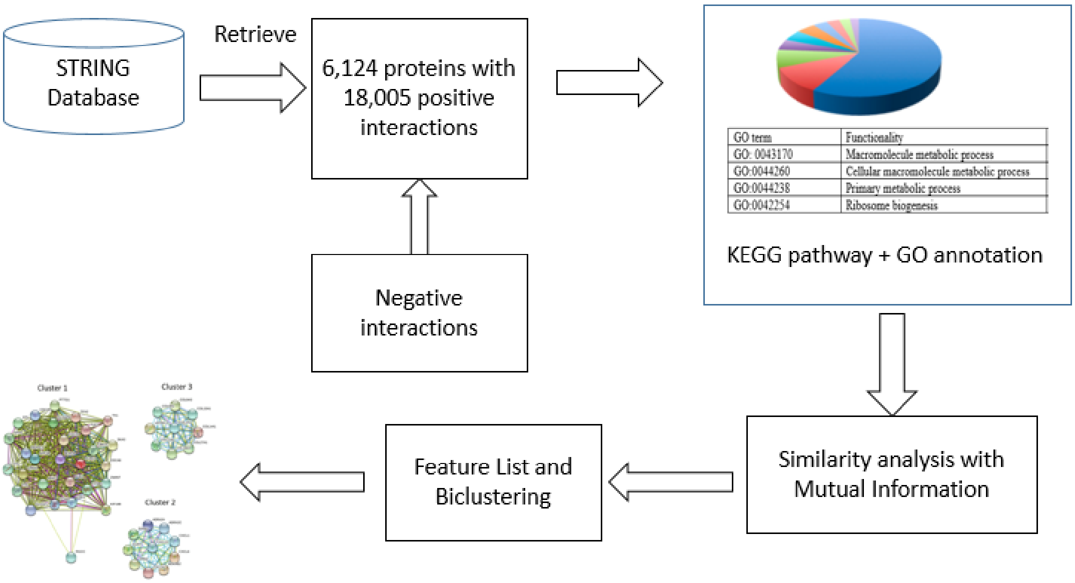
STRING Database
Retrieve
6,124 proteins with 18,005 positive interactions
Negative interactions
GO term
Functionality
GO: 0043170
Macromolecule metabolic process
GO:0044260
Cellular macromolecule metabolic process
GO:0044238
Primary metabolic process
GO:0042254
Ribosome biogenesis
KEGG pathway + GO annotation
Similarity analysis with Mutual Information
Feature List and Biclustering
Cluster 1
Cluster 3
Cluster 2

Fig. 1: System Architecture for Gene Encoding and Biclustering

IV. RESULTS AND DISCUSSION

STRING databases on human protein interactions are taken into consideration for analysis. 12,300 positive connections have been taken into consideration for additional analysis from the protein association file protein.links.detailed.v9.1.txt.gz, which is downloaded [17]. A representative functional connection is shown in Table 1. STRING incorporates a number of sources, including gene fusion, co-expression, conservation in the local area, phylogenetic data, KEGG, Molecular Interaction Database, and BioGRID, for analysing human protein interaction datasets.

Table 1. Functional relationships of protein with 9606 accession code and interaction scores above 0.5.

node1	node2	node1-string-id	node2-string-id	score
AHR	PIN1	9606.ENSP00000242057	9606.ENSP00000425330	0.878
CSNK2A1	ESR1	9606.ENSP00000217244	9606.ENSP00000247170	0.532
ESR1	PIN1	9606.ENSP00000405330	9606.ENSP00000217970	0.481
GLTSCR2	RPL18A	9606.ENSP00000246802	9606.ENSP00000262247	0.766

Analysis is done using interaction scores greater than 0.5, which are regarded as being of high confidence. Combining protein correlations that STRING does not provide results in negative relationships. The dataset was made up of both positive and negative connections. Spreading the negative connections throughout the datasets is crucial since they may have an impact on the outcome. Equations

(1) and (2) show how the 17,931 GO keywords and 271 KEGG pathways from the GO Consortium are represented.

$$GO(p) = [g_1, g_2, \dots g_{17931}] \tag{1}$$

$$Pathway(p) = [k_{1,}\ k_2, \dots . k_{271}] \tag{2}$$

If a protein has a GO word identified, the assignment receives a score of 1. The assignment receives a value of 1 if a KEGG pathway exists, else a score of 0 is given. Each protein was expressed as the result of information from GO terms and pathways [18]. The most pertinent characteristics are picked after comparing each feature's MI to the previously specified protein connections. 172 GO connections in all were chosen, and the chosen relationships were found in most of the datasets. Table 2 lists an example set of associations. QUBIC-R kit is paired with QUBIC version 3.12 to conduct biclustering. Biclusters that are relevant and high-quality are produced using gene expression data.

Table 2. GO associations with MI score above 0.3

GO Association	GO	MI Score
GO:0044260_1+GO:0044260_3	GO:0044260	0.54618
GO:0044446_1+GO:0044446_3	GO:0044446	0.54523
GO:0070013_1+GO:0070013_2	GO:0070013	0.52543
GO:0005634_1+GO:0005634_2	GO:0005634	0.53714

According to the literature, the accuracy of subset selection improves when multiple objectives are considered and it is found to be more optimized. All the GO terms and pathways do not contribute effectively to functional associations [19]. MI scores of the associations in all the 10 datasets are calculated and features with the highest MI are more relevant. QUBIC is executed with all the default parameters like the number of clusters as 100, level of consistency for bicluster or conservation parameter as 0.95, minimum sample size as 2, and quantile discretization as 0.06. The total count of associations with MI score is represented in table 3.

Table 3. The total count of associations with the MI score ranges

Number of Associations	MI score ranges
43	0.3 to 0.4
59	0.4 to 0.5
47	0.5 to 0.6

The Silhouette index is a useful tool for evaluating cluster validity using Euclidean and cosine measurements. It is easier to calculate than the Davies Bouldin index, has clearer interpretation rules, and is more accurate [20]. The Silhouette index measures the cohesion and separation of clusters, with cohesion being the average distance between each point and every other point within the same cluster, and separation being the average distance between every point in one cluster and every other point in the other cluster. A high Silhouette index indicates perfect clustering and is represented in fig 2.

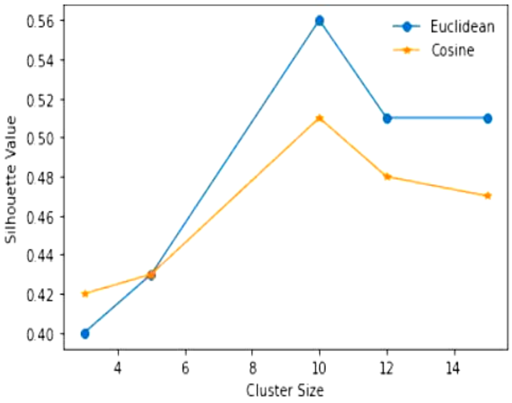

Fig. 2. The performance of Euclidean and Cosine with Silhouette index

If x is a datapoint and a and b stand for cohesion and separation, respectively, then eq (3) for calculation of the silhouette index is used. In all three ontologies, the suggested technique finds the crucial GO words.

$$Silhouette = \frac{(b-a)}{\max(a,b)} \quad (3)$$

Most PPI hub genes are active in cellular and primary metabolic processes, as well as biogenesis. Hence, the proposed method encodes the genes and then bicluster them and it is found to be effective in detecting all the essential hub genes in all three ontologies. The generated bicluster are verified with the Silhouette index. Thus, all the biologically significant GO terms were grouped. Table 4 indicates the essential genes selected under each ontology.

Table 4. Set of genes selected under the ontologies

GO term	Functionality	MI Score
GO:0044428	Nuclear part	0.637
GO: 1901363	Heterocyclic compound	0.622
GO: 0043170	Macromolecule metabolic	0.663
GO:0097159	Organic cyclic binding	0.622
GO:0044260	Cellular macromolecule	0.664

V. CONCLUSION

Analysing gene expression data and understanding biological processes is a crucial undertaking, and the ever-increasing volume of biological data adds to its complexity. Within an interaction network, Gene Ontology (GO) terms play a significant role, serving as influential factors. Building upon this fundamental principle, a model has been developed to identify essential GO terms based on protein functional associations. The proposed approach integrates biological characteristics by means of gene encoding and effectively identifies essential GO terms. Biclusters serve as a representation of the biological interpretation. By clustering genes that are functionally associated under various experimental conditions and utilizing the GO-KEGG encoded matrix, this method offers the advantage of identifying genes with similar functions.

REFERENCES

1. Samish, I., Bourne, P. E., & Najmanovich, R. J. (2015). Achievements and challenges in structural bioinformatics and computational biophysics. *Bioinformatics*, *31*(1), 146-150.
2. Ali, A., Viswanath, R., Patil, S. S., & Venugopal, K. R. (2017, May). A review of aligners for protein protein interaction networks. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)* (pp. 1651-1655). IEEE.
3. Xenarios, I., Fernandez, E., Salwinski, L., Duan, X. J., Thompson, M. J., Marcotte, E. M., & Eisenberg, D. (2001). DIP: the database of interacting proteins: 2001 update. *Nucleic acids research*, *29*(1), 239-241.
4. Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic acids research*, *34*(suppl_1), D535-D539.
5. Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., ... & Jensen, L. J. (2011). Mering Cv. 2011. *The STRING database in*.
6. Ali, A., Hulipalled, V. R., & Patil, S. S. (2020, December). Centrality Measure Analysis on Protein Interaction Networks. In *2020 IEEE International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET)* (pp. 1-5). IEEE.
7. Ali, A., Hulipalled, V. R., Patil, S. S., & Abdulkader, R. (2021). DPEBic: detecting essential proteins in gene expressions using encoding and biclustering algorithm. *Journal of Ambient Intelligence and Humanized Computing*, 1-8.
8. Ahmed, H., Howton, T. C., Sun, Y., Weinberger, N., Belkhadir, Y., & Mukhtar, M. S. (2018). Network biology discovers pathogen contact points in host protein-protein interactomes. *Nature communications*, *9*(1), 2312.
9. Jere, S., Jayannavar, L., Ali, A., & Kulkarni, C. (2017, February). Recruitment graph model for hiring unique competencies using social media mining. In *Proceedings of the 9th International Conference on Machine Learning and Computing* (pp. 461-466).
10. Zhang, X., Xu, J., & Xiao, W. X. (2013). A new method for the discovery of essential proteins. *PloS one*, *8*(3), e58763.
11. Orzechowski, P., Boryczko, K., & Moore, J. H. (2019). Scalable biclustering—the future of big data exploration, *GigaScience*, *8*(7), giz078.
12. Ali, A., Ajil, A., Meenakshi Sundaram, A., & Joseph, N. (2023). Detection of Gene Ontology Clusters Using Biclustering Algorithms. *SN Computer Science*, *4*(3), 217.
13. Cheng, Y., & Church, G. M. (2000, August). Biclustering of expression data. In *Ismb* (Vol. 8, No. 2000, pp. 93-103).
14. Ali, A., Hulipalled, V. R., & Patil, S. S. (2022). A Novel Semantic Similarity Score For Protein Data Analysis. *Computing Technology Research Journal*, *1*(1), 1-4.
15. Patil, S. S., Ali, A., & Ajil, A. (2023). Approaches for Network Analysis in Protein Interaction Network. *International Journal of Human Computations & Intelligence*, *2*(2), 47-54.
16. Li, G., Ma, Q., Tang, H., Paterson, A. H., & Xu, Y. (2009). QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic acids research*, *37*(15), e101-e101.
17. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., ... & Mering, C. V. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, *47*(D1), D607-D613.
18. Ben-Dor, A., Chor, B., Karp, R., & Yakhini, Z. (2002, April). Discovering local structure in gene expression data: the order-preserving submatrix problem. In *Proceedings of the sixth annual international conference on Computational biology* (pp. 49-57).
19. Sathiyamoorthi, V., Ilavarasi, A. K., Murugeswari, K., Ahmed, S. T., Devi, B. A., & Kalipindi, M. (2021). A deep convolutional neural network based computer aided diagnosis system for the prediction of Alzheimer's disease in MRI images. *Measurement*, *171*, 108838.
20. Petrovic, S. (2006, October). A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. In *Proceedings of the 11th Nordic workshop of secure IT systems* (Vol. 2006, pp. 53-64). Citeseer.

21. Kumar, S. S., Ahmed, S. T., Vigneshwaran, P., Sandeep, H., & Singh, H. M. (2021). Two phase cluster validation approach towards measuring cluster quality in unstructured and structured numerical datasets. *Journal of Ambient Intelligence and Humanized Computing*, *12*, 7581-7594.
22. Swamy, R., Ahmed, S. T., Thanuja, K., Ashwini, S., Siddiqha, S., & Fathima, A. (2021, January). Diagnosing the level of Glaucoma from Fundus Image Using Empirical Wavelet Transform. In *Proceedings of the First International Conference on Advanced Scientific Innovation in Science, Engineering and Technology, ICASISET 2020, 16-17 May 2020, Chennai, India*.