Modeling a Gene Structure Behavior Analysis based on the Correlation Ontology

Sudha V, Girijamma H A

Abstract*: The ever increasing digitization and advancement in the medical filed provides data especially related to gene structure and computing models gives an opportunity to analyses those data for the more critical classifications and analysis to provide practitioner a better decision-making platform to advice proper treatment. The subtype classification is a challenging task if it is handled only by the computer vision methods, whereas if the low-level relationship is established and structure of the gene profile is understood then the statistical methods are quite useful and effective for the sub-type doses classifications. This paper presents a process of analyzing the gene structure and its correlations among the node behavior analysis by modeling it at the numerical computing platform. Various performance metrics like p-score and t-test is conducted to get the optimal performance factor. The proposed methods can be extended to the further critical computations in advanced models and get the analysis of typical gene profile structure behaviors and used as an effective classifier for the sub-type classifier of the various type of doses sub-cluster. The computational analysis shows significant improvement (50%) in type-1 and type-2 gene expression analysis.

Keywords: **Biomedical, *Gene Structure*, Gene Ontology, Clustering Support Vector Machine.**

I. INTRODUCTION

In the digital era there as many spaces where a large and highly un-structured data are generated useful in the various domains of applications like social network analysis, medical and businesses. One of the critical area is handling a microarray data for the purpose of understanding gene patterns. The study of the gene requires a very high dimension of feature understanding in their feature ontology. The meaningful understanding of these gene features helps the medical practitioners to decode many of the very valuable insights to develop various thoughtful functions to analyze and diagnosis functions related to biological aspects. There exists a very high level of the sensitiveness into the microarray data as the linking and association among the profile of the expression of the gene is very large scale in the shell of genomics due to the constitutes like cDNA or oligonucleotide. The critical analysis of these large structures of the information plays significant and predominant role into the classification if the accurate statistical analysis is performed, especially to classify among various classes of the disease. There are tremendous numbers of scenarios in the science as well as in the domain of the engineering where we come across a heterogeneous and large amount of data which contains behavioral aspects in variety of perspectives. Traditionally, the mining approach of fused heterogeneous is studied for such smiteful data to evolve methods for large dataset. The approach of handling network to synchronize a meaningful insight is a NP-hard problem. Many researches are conducted on the mapping the node structures for the protein analysis on the ontological context. The technological advancement into the throughput enhancements such techniques includes capturing the conformations of the chromosome that produces even a 3-D structures of the genomes.

The study evolution in the domain of the gene analysis has improvised to the level of the molecular stages by means of gathering the distributed data together. There is an important value of the structures in the protein analysis because it evolves over period of time in contrast to the functional aspects; therefore the computational model shall provide a reliable interpretation because of the annotation constraints. There are various other areas like conceptualization of the drugs, prediction and dysregulations of the protein require accurate transportation of the proteins etc, these methods adopt an expensive experimental method, so a better and efficient computational model is requiring in this direction. In a nutshell, it can be said that a complex structure, high dimension and heterogeneous data insight extraction and predictions can provide many functional attributes to the medical experts to deal with the special conditions. Many critical diseases sub-classification require more critically to be analyzed for any kind of the misjudgment while the diagnosis and this can be achieved by the effective molecule level analysis of the protein. Many methods is evolved such as structural pattern, exchange of information, statistical covariance, network analysis and index based methods.

The subclass classification of particular biomedical disease like cancer is tedious task as cancers are found more than 100 plus of types. The core methodical structures exploration from semantic ontology data is often found computationally challenging task. The prime reason behind this is biomedical ontology structure consist of huge data entities indicating knowledge graph in terms of principles and semantic connectivity. Understanding of the genetic codes corresponding to phenotype outliers is important to distinguish proper disease conditions. Thereby an effective solution model to handle this outlier issue is very much needed from the research and development viewpoint. Analysis of gene expression patterns also reveal more crisp information about the data samples which are crucial to realize the progression of critical disease like cancer sub-classes. The conventional approaches mostly considered limited samples where uncovering the existing facts over time become computationally exhaustive task.

Revised Manuscript Received on December 12, 2019.
***** Correspondence Author
Sudha V*, *Assistant Professor*: Department of IS&E, RNS Institute of Technology, Bengaluru, India, Email: sudhavinayakam@gmail.com
Girijamma H A**, Professor.: Department of Computer Science & Engineering, RNS Institute of Technology, Bengaluru, India

Different computational models emerged to a significant solution level where it considers the statistical features of similarity matrix to explore more about the biomolecular network of protein ingredients.

Another aspect of the research problem is to develop computation model to handle the annotation issue along with problem of signaling. Another importance and open research problem include collaborative gene regulation as today it is possible to get large scale data using higher throughput gene acquisition technologies. The exploitation of evolving larger non coded RNAs can take place by developing efficient models for the annotation of accurate functions. The gene analysis at various layers has got wider scope even into the analysis of HIV suspects, where the accurate prediction of the perturbations plays an important role in order to understand the infection pattern in the protein. The section II explains many of the connected and related work in the direction of gene analysis, section III explains the proposed model for the ontology and its intermediate results and interpretations followed by section IV as a conclusion.

II. LITERATURE SURVEY

An approach of fusion of the data of various variety is studied by (M. Žitnik and B. Zupan, 2015) for establishing the relation among the context by using the concept of factorization for various different data sources of gene and exhibits better accuracy as compared to the analysis on the single set of data[1].The synchronization of the biological node network is proposed by (J. Xie *et al*.,2016) by a hybrid approach by combining two classical algorithm Hungarian along with the greedy one. Where a pair of the nodal structure is mapped for the protein analysis to get an isolated ontology of the gene with an objective function of minimizing the entropy [2].The use of the non-negative factorization is exploited by the (X. Li and K. Wong et al , 2018) for the purpose of the localization of the cluster at the local space by an optimization technique by validating it on the chromosome maps for reduced time complexity to obtain higher accuracy of classification[3].

The specific condition of the disease classification is studied by (Liu et al, 2018) by means of analysis of the molecular structure and it's interaction on the basis of the semantics and the pattern of the sequence[4]. The protein analysis and synthesis reveal the functional interactions and regulations among molecular protein agents and can be realized through structural patterns. (Yerneni et al, 2018) introduced a robust schema to effectively determine the interactions and exchange of information between two protein agents [5]. The model considers a substantial scoring method in order to predict the associativity metric computed on the basis of Gene based numerical analysis. (Piras et al, 2019) explored the regulation of the pattern of genotypes using a statistical covariance estimation process [6]. The procedure is functionally mechanized to use biofilms to bring more insight into genomic structural expressions. The complex mode of interactions in protein-protein networked structures has been investigated through a methodical procedure by (He and Chan, 2019). A index-based metric is used to measure the associativity score which also indicates the correlation matrix between two protein structural and molecular agents [7]. The validation process shows the effectiveness. The extensive analysis of variable expression-oriented gene structure is realized on the basis of gene ontology by (Vehdat and Bakhshandeh, 2019). The efficient profiling of cell entities revealed significant improvement in the detection process of chemical signalling route patterns in cell structures [8].

In the work proposed by (J. Si, X. Zhao, X. Zhao and R. Wu, 2015), a method adopted by the bioinformatics technology for the large library of the cDNA data to identified differentially presented structure for achieving the visual representations of the gene structure or the network[9]. The annotation problem to handle the dynamic behaviors of the structure is handled in the work of (L. Deng and Z. Chen, 2015) by introducing a predictive mechanism of efficient computational model to handle the verity of the structure to combat the heterogeneity using the Bayesian network. The effectiveness of this model is studied by validating its performance with the uni-functionally dataset of the protein structure by measuring its F-score and coverage aspects [10]. The approach of machine learning using Support vector machine (SVM) is used by (L. Li *et al2016) in their work for the purpose of the predicting the transportation mechanism of the membrane proteins and validates their model of the prediction using the cross-validation tests namely* jackknife to analysis the accuracy of prediction [11]. One another work towards the membrane protein by (S. Wan, M. Mak and S. Kung, 2016) works over improvements of the interpretability limitations of the conventional methods by proposing a new dimension of prediction with the sparse and high interpretation among the ontology of the gene using elastic net classifier [12].(M. Masseroli, A. Canakoglu and S. Ceri, 2016) created an architecture of solution strategy as a tool to manage the knowledge reposit of the genomic information to maintain the quality along with the consistency to explore the contextual ontology based relationship analysis[13].

Proper identification of defected gene-patterns from different cancerous modular structure is crucial and the exploration of beneficial targets is important. The authors (R. K. Makhijani, S. A. Raut, H. J. Purohit, 2018), computationally explored the gene expression patterns (i.e. *RNA protein seq*) from micro-array structural tool to properly bring more insight into molecular interrelations. An exhaustive statistical modeling set is incorporated to analyse the extracted gene data expressions. A connection matrix also indicates the connection vectors between principle protein agents which are more important to attain gene regulatory functional specifications [14]. (Q. Zhang and D. Haglin, 2016) attempted to provide a methodology capable of exploiting the biomedical ontology pattern from gene expression data. The prime motive is to ensure disclosure of different levels of semantic similarity. The study addresses the computational complexity problem of ontology pattern exploration and provides methodical solution schema to significantly reduce the size of extracted gene data without compromising the deeper substantial insight [15]. (M. Glueck et al, 2017) introduced a computational approach which is analytics assisted and perform classification of phenotype patterns on the basis of ontology clustering policy. It also effectively reveals the ontology topological structure to get more insight into error-prone quality information [16].

The authors (M. Kim, D. Kim and J. Kim, 2019), considered cancer related gene expression data profiles and developed a model based on regression analysis to verify the stage of cancer development. It also explored the sub-gene network type to bring more critical information corresponding to underlying metastasis [17]. (Y. Shui and Y. Cho, 2016) incorporated a robust modeling of semantic gene ontology to explore the potential features of protein complexes in order to learn more about the connection vector. The model exploits the semantic features to measure the similarity index and it basically assist in prediction of complex molecular structures [18].

The issue of annotation in the work of (G. Yu, G. Fu, J. Wang and Y. Zhao, 2018), I s handled by a predictive model for the annotation ontology using a graph method for encoding [19]. The issue of the networks signalling is widely studied by (Y. Cho, Y. Xin and G. Speegle, 2015) by presenting the network of signal using directed acyclic graph (DAG) as well developed a web-based tool for the visualization of the signaling network [20].The collaborative gene regulation is studied by (J. Luo, G. Xiang and C. Pan, 2017) by using a optimized method of matrix factorization to combine varied expression profile and introduces a penalty factor to minimize sparsity using mathematical model[21]. The use of larger non coded RNAs for prediction of important function is represented as a network in the work of (Z. Zhang, J. Zhang, C. Fan, Y. Tang and L. Deng, 2019), for three different and heterogenous networks by computation of the similarity and its effectiveness is validated against the conventional methods of that time using F-score [22]. The authors (S. Ray and U. Maulik, 2018) contributed for analyzing the HIV infection by proposing a method of gene differentiation on selective feature set using matrix factorization. The performance is checked by the co-regulation symptom [23]

Table.1 Summarization of few literatures

Literatures	Approach
[19]	Predictive modeling for annotation ontology assessment
[20]	Approach to visualize signaling network
[21]	Minimizing sparsity based on mathematical approach
[22]	Predictive model for important functions of proteins behaviour
[23]	Method to analysis HIV infection
[24]	Assessment of gene expression data

III. RESULTS AND DISCUSSIONS

The study designed an analytical research methodology considering microarray data modular structure in order to analyze gene expression data from the classification viewpoint. The design modeling of is improvised satisfying the operational constraints to enhance the structural pattern analysis of gene expressions in proteins.

An architectural design of the framework for analyzing the gene data for various clinical advantages using the gene perspective from the data store of the microarray to interpretation at the significant functionality's levels. The micro-array (M_A) contains a set of expressions-oriented pattern associated with protein structural molecules of cDNA. It is further compared with the gene expression profiles within a defined range of scale R_{scale} . There is a need to develop a procedural technique through which the differences in gene expression can be statistically measured to realize different conditions of nucleic acids chromosomal status. The study considered gene expression data set related to different types of tumor conditions of embryonic cell structures [24].

- **Dataset:** Refer dataset used in [24]

The framework takes up gene expression datastore containing three significant patterns of RMA, GCRMA with MLE and GCRMA with EBE, with Number of gene (N) and number of samples(S). The corelated structure among thcorresponding gene symbols for each gene (N) with the respective probes is taken up to annotate the values of the expression with the respective symbols. A small set n € N is tabulated in the table 1 technical reason initiates data missing or impure data stage, those are filtered and as an identifier are replaced by a unique symbol using a low-value filter process and a smaller profile variance filter, which reduces the value of N→ N' s.t N' <N.

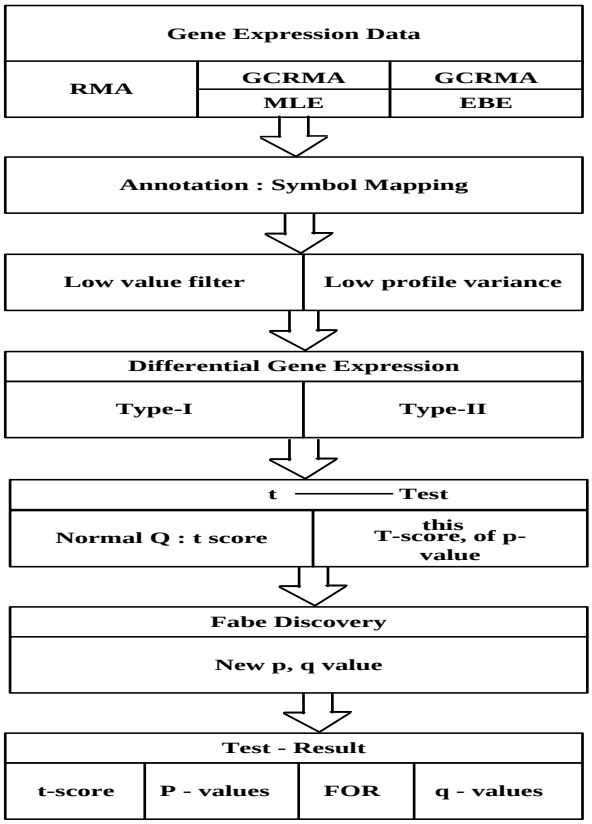

Fig.1 Block-based flow of the formulated procedure

The corresponding multiclass level is set as a data index in the form of matrix. The system failures and another The further process includes considering all the 'S' sample and k€ N' dataset from type-1 and type -2 group of gene expression for the purpose of comparison, where the comparison takes place on the basis of label annotation of respective type-1 and type-2.

In order to understand the predominate morphing in the expression of type1 and type2 samples; a t-test is performed. The figure 1 and 2 respectively.

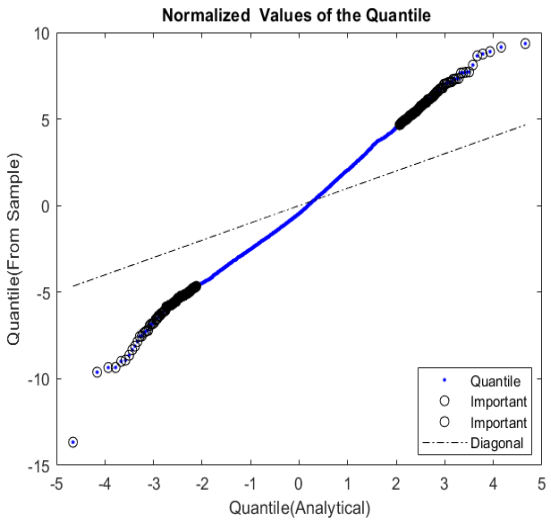

Fig.2 Normal quintile plots of t-scores

Figure 2 exhibits the outcome obtained after performing numerical analysis on predominate morphing on the sub cluster of type-1 and type-2 samples of gene expression data. It shows the normalized quintile values along with significant data points of particular gene expression.

The analysis of t-score and p-value evaluation is performed in a numerical computing platfom This is performed in the gene expression data samples in order to identify the significicant changes in the pattern of genes. This identification of significant changes in the gene can be negative or positive in when concerned from the diaqgnostic viewpoint in gene data exprssion analysis.

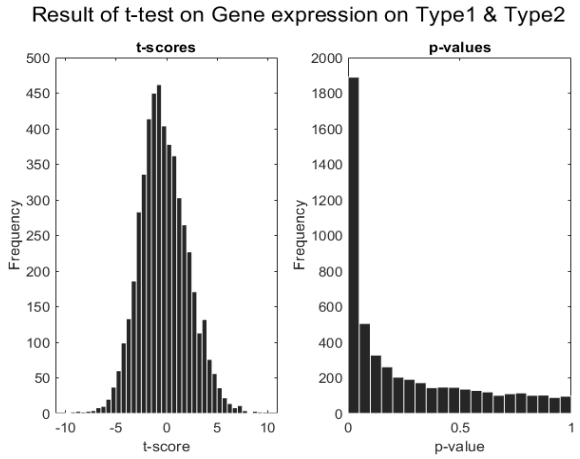

Fig.3 histograms of t-scores and *p-values* of the t-tests

Table.2 Quantitative analysis of t-scores

t-scores	-10	-5	0	5	10
Frequency	0.1	52.3	452	48	0.1

Table.3 Quantitative analysis of p-values

p-values	0	0.5	1
Frequency	1821	180	160

The figure 3 shows the outcome of the extensive numerical analysis on the test sample values of type-1 and type-2 gene expression data. It also shows the frequency of t-scores and p-values obtained after performing t-test on the sub-clusters of two different types of gene expression data. The prime objective here is to evaluate the gene expression to determine the phenotypes or functional conditions among proteins.

Table.4 A small subset of gene data sample

DCK	TGFB3
HIST1H2AE	EVI2B
FPR1	HLX1
RPS16	NCL
RPS6KB1	HIST1H1E
LOT2'	…N

Table.5 Significant features and their overview form gene data sample

Gene expression representation features	**Dimension**
ExpressionType1(cns/gcrma/eb)	(6x1)
ExpressionType2(cns/gcrma/mle)	(6x1)
ExpresionType3(cns/rma)	(6x1)
Gene symbol map: [Count KeyType ValueType]	

The procedure also estimates binomial coefficient considering a sample size of different gene expression G_{ex}.

$$\forall G_{ex} \in N, S\ s.t.\ G_{ex} \rightarrow Z^{+} \quad \text{eq. 1}$$

The computation of binomial coefficients to evaluate the structural pattern of gene expression can be mathematically expressed as following eq. 2

$$G_{ex}C_i = \binom{G_{ex}}{i} \leftarrow \frac{G_{ex}!}{(G_{ex-}i)!\,G_{ex}!} \quad \text{eq. 2}$$

The structural analysis of gene expression data reveals the fact that in during analysis the procedure can encounter two different types of error i.e. false positive error, it indicates that a gene is differentially expressed but in reality it is not and on the other hand, the procedure can also come across false negative error which indicates that the test fails to identify the actual pattern of gene expression and its regulation functions. The procedure incorporated Storey-Trbshirani method to compute the q-value and p-value test. This test also ensures positive false discovery rate which can strengthen the analysis to identify the actual gene expression pattern with a proper classification measure [25]-[28].

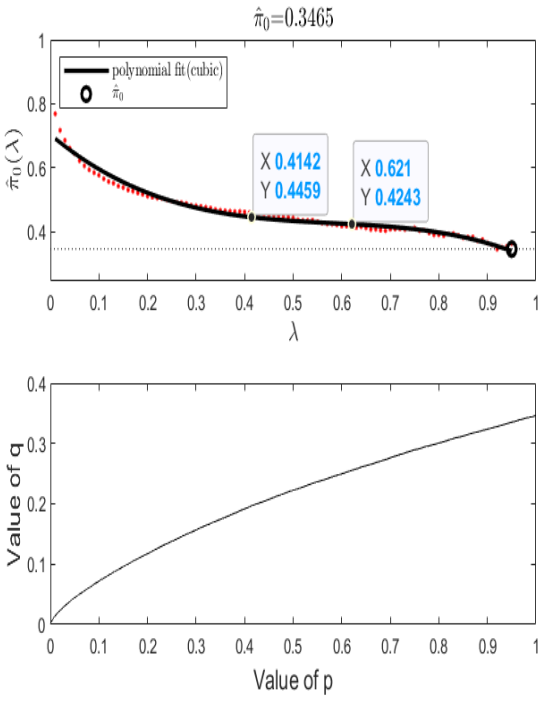

Fig.4 q-value estimation considering null-hypothesis

The procedure also incorporates a study of null-hypothesis of thousands of genes to measures the minimum false discovery rate that occurs during the test procedure execution.
The following figure shows the architectural block-based diagram which shows the procedure to assess the gene expression analysis as adopted in this research study.

Figure 4 shows the flow design that is meant to evaluate the analysis for identifying the gene expression data. It shows that in the initial phase of computation the procedure import the gene expression data from the data base and further apply annotation and symbol mapping paradigm to filter out low value filter and low profile variance factors. After this the estimation of two different types of gene expression is evaluated considering sub-clustering analysis. Further the experimental analysis considers t-test in terms of normal Q-t score and p-value analysis. Finally, the test results are evaluated statistically significant features of GO. In this phase the effective evaluation of annotated gene and the number of up-regulatory genes play a very crucial role. The study shows that the analysis is computationally efficient and provides better insight into gene protein structures in terms of proper clustering.

IV. CONCLUSION AND FUTURE RESEARCH

The underlying concept of data-clustering has significantly gained more attention from the research community for its potential capability to explore the hidden pattern of inherent structure of data. The structural sequence of data can be divided into disjoint or overlapping groups for the computation based operational needs. However, owing to popular advantages, the data clustering problem emerged and narrowed down in the context of Gene expression data analysis to assist in better classification of disease conditions. As it poses higher scope of applicability into medical research and identification of critical disease conditions, thereby it has become an active research area. Existing microarray tool-based analysis and methodical procedures poses a great deal of challenges when the extraction of Gene data expressions is concerned. The prime reason behind this is that- the extraction process often introduces artifacts which make the diagnosis and classification of data computationally challenging and intensive. Thereby, to balance the computational and accuracy performance of medical disease diagnosis procedure especially in the case of cancer detection, the classification mechanism has to be well-furnished and less-iterative. This study addresses this problem and introduces a robust computational approach which assists in complex gene data expression analysis and classification using microarray-gene expression database. The extensive numerical analysis is modeled to justify the outcome eventually. This effective analysis has a higher scope of applicability into proper/accurate diagnosis of diseases like cancer from computational viewpoint.

REFERENCES

1. M. Žitnik and B. Zupan, "Data Fusion by Matrix Factorization," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 1, pp. 41-53, 1 Jan. 2015.
2. J. Xie et al., "An Adaptive Hybrid Algorithm for Global Network Alignment," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no. 3, pp. 483-493, 1 May-June 2016.doi: 10.1109/TCBB.2015.2465957
3. X. Li and K. Wong, "A Comparative Study for Identifying the Chromosome-Wide Spatial Clusters from High-Throughput Chromatin Conformation Capture Data," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 15, no. 3, pp. 774-787, 1 May-June 2018.doi: 10.1109/TCBB.2017.2684800
4. Y. Liu et al., "Prediction of cancer-associated piRNA–mRNA and piRNA–lncRNA interactions by integrated analysis of expression and sequence data," in Tsinghua Science and Technology, vol. 23, no. 2, pp. 115-125, April 2018.doi: 10.26599/TST.2018.9010056
5. S. Yerneni, I. K. Khan, Q. Wei and D. Kihara, "IAS: Interaction Specific GO Term Associations for Predicting Protein-Protein Interaction Networks," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 15, no. 4, pp. 1247-1258, 1 July-Aug. 2018.
6. V. Piras, A. Chiow and K. Selvarajoo, "Long-range order and short-range disorder in Saccharomyces cerevisiae biofilm," in Engineering Biology, vol. 3, no. 1, pp. 12-19, 3 2019.doi: 10.1049/enb.2018.5008
7. T. He and K. C. C. Chan, "Measuring Boundedness for Protein Complex Identification in PPI Networks," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 3, pp. 967-979, 1 May-June 2019.doi: 10.1109/TCBB.2018.2822709
8. S. Vahdat and B. Bakhshandeh, "Prediction of putative small molecules for manipulation of enriched signalling pathways in hESC-derived early cardiovascular progenitors by bioinformatics analysis," in IET Systems Biology, vol. 13, no. 2, pp. 77-83, 4 2019.
9. [9] J. Si, X. Zhao, X. Zhao and R. Wu, "Systematic functional genomics resource and annotation for poplar," in IET Systems Biology, vol. 9, no. 4, pp. 164-171, 8 2015.
10. L. Deng and Z. Chen, "An Integrated Framework for Functional Annotation of Protein Structural Domains," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 12, no. 4, pp. 902-913, 1 July-Aug. 2015.
11. L. Li et al., "Prediction the Substrate Specificities of Membrane Transport Proteins Based on Support Vector Machine and Hybrid Features," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no. 5, pp. 947-953, 1 September 2016. doi: 10.1109/TCBB.2015.2495140
12. S. Wan, M. Mak and S. Kung, "Mem-mEN: Predicting Multi-Functional Types of Membrane Proteins by Interpretable Elastic Nets," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no. 4, pp. 706-718, 1 July-Aug. 2016. doi: 10.1109/TCBB.2015.2474407

13. M. Masseroli, A. Canakoglu and S. Ceri, "Integration and Querying of Genomic and Proteomic Semantic Annotations for Biomedical Knowledge Extraction," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no. 2, pp. 209-219, 1 March-April 2016.
14. R. K. Makhijani, S. A. Raut and H. J. Purohit, "Fold change based approach for identification of significant network markers in breast, lung and prostate cancer," in IET Systems Biology, vol. 12, no. 5, pp. 213-218, 10 2018.doi: 10.1049/iet-syb.2018.0012
15. Q. Zhang and D. Haglin, "Semantic similarity between ontologies at different scales," in IEEE/CAA Journal of AutomaticaSinica, vol. 3, no. 2, pp. 132-140, 10 April 2016.doi: 10.1109/JAS.2016.7451100
16. M.Glueck, A. Gvozdik, F. Chevalier, A. Khan, M. Brudno and D. Wigdor, "PhenoStacks: Cross-Sectional Cohort Phenotype Comparison Visualizations," in IEEE Transactions on Visualization and Computer Graphics, vol. 23, no. 1, pp. 191-200, Jan. 2017. doi: 10.1109/TVCG.2016.2598469
17. M. Kim, D. Kim and J. Kim, "Stage-Dependent Gene Expression Profiling in Colorectal Cancer," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 5, pp. 1685-1692, 1 Sept.-Oct. 2019.doi: 10.1109/TCBB.2018.2814043
18. Y. Shui and Y. Cho, "Alignment of PPI Networks Using Semantic Similarity for Conserved Protein Complex Prediction," in IEEE Transactions on NanoBioscience, vol. 15, no. 4, pp. 380-389, June 2016.
19.]G. Yu, G. Fu, J. Wang and Y. Zhao, "NewGOA: Predicting New GO Annotations of Proteins by Bi-Random Walks on a Hybrid Graph," in IEEE/ACM Transactions on ComputationalBiology and Bioinformatics, vol. 15, no. 4, pp. 1390-1402, 1 July-Aug. 2018.
20. Z. Zhang, J. Zhang, C. Fan, Y. Tang and L. Deng,"KATZLGO: Large-Scale Prediction of LncRNA Functions by Using the KATZ Measure Based on Multiple Networks," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 2, pp. 407-416, 1 March-April 2019.
21. Y. Cho, Y. Xin and G. Speegle, "P-Finder: Reconstruction of Signaling Networks from Protein-Protein Interactions and GO Annotations," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 12, no. 2, pp. 309-321, March-April 2015.
22. J. Luo, G. Xiang and C. Pan, "Discovery of microRNAs and Transcription Factors Co-Regulatory Modules by Integrating Multiple Types of Genomic Data," in IEEE Transactions on NanoBioscience, vol. 16, no. 1, pp. 51-59, Jan. 2017
23. S. Ray and U. Maulik, "Discovering Perturbation of Modular Structure in HIV Progression by Integrating Multiple Data Sources Through Non-Negative Matrix Factorization," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 15, no. 3, pp. 869-877, 1 May-June 2018.
24. Pomeroy, S.L., et al., "Prediction of central nervous system embryonal tumour outcome based on gene expression". Nature, 415(6870):436-42, 2001
25. Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. Royal Stat. Soc. 57:289–300.
26. Storey, J.D., and Tibshirani, R. 2003. Statistical significance for genomewide studies. Proc. Nat. Acad. Sci. 100:9440–9445.
27. Storey, J.D., Taylor, J.E., and Siegmund, D. 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. J. Royal Stat. Soc. 66:187–205.
28. Storey, J.D. 2002. A direct approach to false discovery rates. J. Royal Stat. Soc. 64:479–498.

AUTHORS PROFILE

Sudha V, Currently working as a Assistant Professor in Information Science and Engineering in RNS Institute of Technology, Bangalore and having a teaching experience of 13 years. Her areas of interests are in the field of Data mining, Data analytics and Machine learning algorithms.

Dr. Girijamma H A, Currently working as a Professor in the department of Computer Science and Engineering in RNS Institute of Technology, Bangalore and having experience of 26 years in teaching. Her research interests are in the areas of automata, compilers, fuzzy logic, data analytics, software engineering, image processing, natural language processing and Machine learning.

.