A Group Average Cluster Analysis of Few IGF1R Sequences using Modified Group Average Link Clustering Algorithm

R. Rambabu
Associate Professor & Head of the Department,
Department of Computer Science & Engineering,
Rajamahendri Institute of Engineering & Technology, Rajahmundry, 533107

P. Srinivasa Rao, PhD
Professor, Department of Computer Science & Systems Engineering,
College of Engineering,
Andhra University,
Visakhapatnam, 530003

ABSTRACT
Clustering techniques have been widely used in the fields of information technology, biomedical sciences. Cluster analysis deals with the identification of a set of objects into subsets with some sort of similarities. Such groups are assigned to have similar function. In this paper, a modified group average clustering program was written in python language and applied on a dataset of IGF1R protein sequences to generate orthologous clusters of sequences and the phylogenetic trees were presented.

Keywords
IGF1R, clusters, group average clustering, python program.

1. INTRODUCTION
Cluster analysis is generally implemented to find partitioning of objects in a given dataset into clusters which suggests that the data points that appear within one group are regarded more similar to one another than the data points that apparently appear in other clusters. Considering the gene expression data, several clustering algorithms have been put forward by scientific community, however, modifications and/or new algorithms have also been proposed. These algorithms have demonstrated to be beneficial in clustering groups of genes and samples with relevance to biology [1] [2].

Partitioning techniques such as Clustering methods might assist us in understanding functional aspects of gene and its regulation, processes that takes place within a cell etc. Genes with similar expression patterns (co-expressed genes) can be clustered together with similar cellular functions. This sort of methodology would aid in understanding the functional relationships among many genes [3] [4].

Clustering is an example of unsupervised classification. Clustering analysis is distinguished from pattern recognition techniques such as discriminant and decision analysis, which find rules for categorizing objects from a given set of pre-classified objects. Cluster analysis would provide insights regarding data distribution.

Compared to other methods of clustering, hierarchical clustering produces a hierarchical series of nested clusters which can be graphically represented by a tree, called dendrogram. Hierarchical clustering algorithms can be further divided into agglomerative and divisive [5] [6] [7] [8] [9].

Hierarchical clustering not only groups together genes with similar expression pattern but also provides a natural way to graphically represent the data set. An agglomerative algorithm called UPGMA (Unweighted Pair Group Method with Arithmetic Mean) was adopted in this paper to study phylogenetic relationships among IGF1R sequences. Insulin-like growth factor 1 receptor (IGF1R) is a transmembrane tyrosine kinase that is widely found in many cell types. IGF1R is a regulator that is vital to growth, differentiation and apoptosis. IGF1R-mediated signaling is crucial for the development and progression of multiple types of cancer.

2. MATERIALS AND METHODS
2.1 Data Set
The orthologous protein sequences selected in this study was extracted from Swiss-Prot database [10]. Criteria implemented to download IGF only sequences by using gene_name:IGF1R tag to eliminate other substrate sequences and the like. After removing shorter length sequences manually by visual inspection, the final data set of 6 sequences are extracted in fasta formats (given below) and subjected to multiple sequence analysis using clustalw [11]. The summary file output of clustalw contains all the necessary data that is required to perform clustering. The pairwise scores generated by the program are given as input to the python [12] based group average clustering algorithm and the phylogenetic trees are reported [13] [14].

```
>sp|P08069|IGF1R_HUMAN Insulin-like growth factor 1 receptor OS=Homo sapiens GN=IGF1R PE=1 SV=1
MKSGSGGGSPTSLWGLLFLSAALSLWPTSGEICGPGIDIRNDYQQLKRLENCTVIEGYLH
ILLISKAEDYRSYRFPKLTVITEYLLLFRVAGLESLGDLFPNLTVIRGWKLFYNYALVIF
EMTNLKDIGLYNLRNITRGAIRIEKNADLCYLSTVDWSLILDAVSNNYIVGNKPPKECGD
LCPGTMEEKPMCEKTTINNEYNYRCWTTNRCQKMCPSTCGKRACTENNECCHPECLGSCS
APDNDTACVACRHYYYAGVCVPACPPNTYRFEGWRCVDRDFCANILSAESSDSEGFVIHD
GECMQECPSGFIRNGSQSMYCIPCEGPCPKVCEEEKKTKTIDSVTSAQMLQGCTIFKGNL
LINIRRGNNIASELENFMGLIEVVTGYVKIRHSHALVSLSFLKNLRLILGEEQLEGNYSF
YVLDNQNLQQLWDWDHRNLTIKAGKMYFAFNPKLCVSEIYRMEEVTGTKGRQSKGDINTR
NNGERASCESDVLHFTSTTTSKNRIIITWHRYRPPDYRDLISFTVYYKEAPFKNVTEYDG
QDACGSNSWNMVDVDLPPNKDVEPGILLHGLKPWTQYAVYVKAVTLTMVENDHIRGAKSE
ILYIRTNASVPSIPLDVLSASNSSSQLIVKWNPPSLPNGNLSYYIVRWQRQPQDGYLYRH
NYCSKDKIPIRKYADGTIDIEEVTENPKTEVCGGEKGPCCACPKTEAEKQAEKEEAEYRK
VFENFLHNSIFVPRPERKRRDVMQVANTTMSSRSRNTTAADTYNITDPEELETEYPFFES
```

```
RVDNKERTVISNLRPFTLYRIDIHSCNHEAEKLGCSASNFVFARTMPAEGADDIPGPVTW
EPRPENSIFLKWPEPENPNGLILMYEIKYGSQVEDQRECVSRQEYRKYGGAKLNRLNPGN
YTARIQATSLSGNGSWTDPVFFYVQAKTGYENFIHLIIALPVAVLLIVGGLVIMLYVFHR
KRNNSRLGNGVLYASVNPEYFSAADVYVPDEWEVAREKITMSRELGQGSFGMVYEGVAKG
VVKDEPETRVAIKTVNEAASMRERIEFLNEASVMKEFNCHHVVRLLGVVSQGQPTLVIME
LMTRGDLKSYLRSLRPEMENNPVLAPPSLSKMIQMAGEIADGMAYLNANKFVHRDLAARN
CMVAEDFTVKIGDFGMTRDIYETDYYRKGGKGLLPVRWMSPESLKDGVFTTYSDVWSFGV
VLWEIATLAEQPYQGLSNEQVLRFVMEGGLLDKPDNCPDMLFELMRMCWQYNPKMRPSFL
EIISSIKEEMEPGFREVSFYYSEENKLPEPEELDLEPENMESVPLDPSASSSSLPLPDRH
SGHKAENGPGPGVLVLRASFDERQPYAHMNGGRKNERALPLPQSSTC

>sp|P24062|IGF1R_RAT Insulin-like growth factor 1 receptor OS=Rattus norvegicus GN=Igf1r PE=2 SV=2
MKSGSGGGSPTSLWGLVFLSAALSLWPTSGEICGPGIDIRNDYQQLKRLENCTVIEGFLH
ILLISKAEDYRSYRFPKLTVITEYLLLFRVAGLESLGDLFPNLTVIRGWKLFYNYALVIF
EMTNLKDIGLYNLRNITRGAIRIEKNADLCYLSTIDWSLILDAVSNNYIVGNKPPKECGD
LCPGTLEEKPMCEKTTINNEYNYRCWTTNRCQKMCPSVCGKRACTENNECCHPECLGSCH
TPDDNTTCVACRHYYYKGVCVPACPPGTYRFEGWRCVDRDFCANIPNAESSDSDGFVIHD
GECMQECPSGFIRNSTQSMYCIPCEGPCPKVCGDEEKKTKTIDSVTSAQMLQGCTILKGN
LLINIRRGNNIASELENFMGLIEVVTGYVKIRHSHALVSLSFLKNLRLILGEEQLEGNYS
FYVLDNQNLQQLWDWNHRNLTVRSGKMYFAFNPKLCVSEIYRMEEVTGTKGRQSKGDINT
RNNGERASCESDVLRFTSTTTWKNRIIITWHRYRPPDYRDLISFTVYYKEAPFKNVTEYD
GQDACGSNSWNMVDVDLPPNKEGEPGILLHGLKPWTQYAVYVKAVTLTMVENDHIRGAKS
EILYIRTNASVPSIPLDVLSASNSSSQLIVKWNPPTLPNGNLSYYIVRWQRQPQDGYLFR
HNYCSKDKIPIRKYADGTIDVEEVTENPKTEVCGGDKGPCCACPKTEAEKQAEKEEAEYR
KVFENFLHNSIFVPRPERRRRDVLQVANTTMSSRSRNTTVADTYNITDPEEFETEYPFFE
SRVDNKERTVISNLRPFTLYRIDIHSCNHEAEKLGCSASNFVFARTMPAEGADDIPGPVT
WEPRPENSIFLKWPEPENPNGLILMYEIKYGSQVEDQRECVSRQEYRKYGGAKLNRLNPG
NYTARIQATSLSGNGSWTDPVFFYVPAKTTYENFMHLIIALPVAILLIVGGLVIMLYVFH
RKRNNSRLGNGVLYASVNPEYFSAADVYVPDEWEVAREKITMNRELGQGSFGMVYEGVAK
GVVKDEPETRVAIKTVNEAASMRERIEFLNEASVMKEFNCHHVVRLLGVVSQGQPTLVIM
ELMTRGDLKSYLRSLRPEVENNLVLIPPSLSKMIQMAGEIADGMAYLNANKFVHRDLAAR
NCMVAEDFTVKIGDFGMTRDIYETDYYRKGGKGLLPVRWMSPESLKDGVFTTHSDVWSFG
VVLWEIATLAEQPYQGLSNEQVLRFVMEGGLLDKPDNCPDMLFELMRMCWQYNPKMRPSF
LEIIGSIKDEMEPSFQEVSFYYSEENKPPEPEELEMELELEPENMESVPLDPSASSASLP
LPERHSGHKAENGPGVLVLRASFDERQPYAHMNGGRANERALPLPQSSTC

>sp|Q60751|IGF1R_MOUSE Insulin-like growth factor 1 receptor OS=Mus musculus GN=Igf1r PE=1 SV=3
MKSGSGGGSPTSLWGLVFLSAALSLWPTSGEICGPGIDIRNDYQQLKRLENCTVIEGFLH
ILLISKAEDYRSYRFPKLTVITEYLLLFRVAGLESLGDLFPNLTVIRGWKLFYNYALVIF
EMTNLKDIGLYNLRNITRGAIRIEKNADLCYLSTIDWSLILDAVSNNYIVGNKPPKECGD
LCPGTLEEKPMCEKTTINNEYNYRCWTTNRCQKMCPSVCGKRACTENNECCHPECLGSCH
TPDDNTTCVACRHYYYKGVCVPACPPGTYRFEGWRCVDRDFCANIPNAESSDSDGFVIHD
DECMQECPSGFIRNSTQSMYCIPCEGPCPKVCGDEEKKTKTIDSVTSAQMLQGCTILKGN
LLINIRRGNNIASELENFMGLIEVVTGYVKIRHSHALVSLSFLKNLRLILGEEQLEGNYS
FYVLDNQNLQQLWDWNHRNLTVRSGKMYFAFNPKLCVSEIYRMEEVTGTKGRQSKGDINT
RNNGERASCESDVLRFTSTTTWKNRIIITWHRYRPPDYRDLISFTVYYKEAPFKNVTEYD
GQDACGSNSWNMVDVDLPPNKEGEPGILLHGLKPWTQYAVYVKAVTLTMVENDHIRGAKS
EILYIRTNASVPSIPLDVLSASNSSSQLIVKWNPPTLPNGNLSYYIVRWQRQPQDGYLYR
HNYCSKDKIPIRKYADGTIDVEEVTENPKTEVCGGDKGPCCACPKTEAEKQAEKEEAEYR
KVFENFLHNSIFVPRPERRRRDVMQVANTTMSSRSRNTTVADTYNITDPEEFETEYPFFE
SRVDNKERTVISNLRPFTLYRIDIHSCNHEAEKLGCSASNFVFARTMPAEGADDIPGPVT
WEPRPENSIFLKWPEPENPNGLILMYEIKYGSQVEDQRECVSRQEYRKYGGAKLNRLNPG
NYTARIQATSLSGNGSWTDPVFFYVPAKTTYENFMHLIIALPVAILLIVGGLVIMLYVFH
RKRNNSRLGNGVLYASVNPEYFSAADVYVPDEWEVAREKITMNRELGQGSFGMVYEGVAK
GVVKDEPETRVAIKTVNEAASMRERIEFLNEASVMKEFNCHHVVRLLGVVSQGQPTLVIM
ELMTRGDLKSYLRSLRPEVEQNNLVLIPPSLSKMIQMAGEIADGMAYLNANKFVHRDLAA
RNCMVAEDFTVKIGDFGMTRDIYETDYYRKGGKGLLPVRWMSPESLKDGVFTTHSDVWSF
GVVLWEIATLAEQPYQGLSNEQVLRFVMEGGLLDKPDNCPDMLFELMRMCWQYNPKMRPS
FLEIIGSIKDEMEPSFQEVSFYYSEENKPPEPEELEMELEMEPENMESVPLDPSASSASL
PLPERHSGHKAENGPGPGVLVLRASFDERQPYAHMNGGRANERALPLPQSSTC

>sp|O73798|IGF1R_XENLA Insulin-like growth factor 1 receptor OS=Xenopus laevis GN=igf1r PE=1 SV=1
MKAELVPVCTAWILGLLLCLGPAAAKVCGPNMDIRNDVSELKQLRDCVVIEGYLQILLIS
NAKAEDFRNLRFPNLTVITDYLLLFRVSGLVSLSNLFPNLTVIRGRVLFYNYALVIFEMT
DLKEIGLYNLRNITRGAVRIEKNSELCYVSTVDWSLVLDAVYNNYIVGNKPPKECVDLCP
GAREKMQICEKSSINNEFADRCWSDEHCQKVCPSVCGKRACSDNNECCHPECLGSCTAPD
NDTACVACHHYFYEGRCVPTCPSNTYKFEGWRCITREVCAKMHIWIHSTIPFIIHKGECV
YECPSGYMLNKSQSMTCSPCEGPCPKICEEKMKTIDSVTSAQMLEGCTVLKGNLQLNIRK
GQNIAAELENFLGLIETVTGYVKIRHSHALVSLSFLKSLRYILGEEQMPGNYSFYVFDNN
NLQQLWDWSKHNLTIKEGKIRFAFNSKLCASEIYRMEEVTGTKGRQAEEDISLSTNGNMA
SCESHVLNFTSRSKIKNRIKLTWERYRPPDYRDLISFTVYYKEAPFRNVTEYDGQDACGS
NSWNMVDVDLPASKESDPGILLQGLKPWTQYAIYVKAITLTMLENRHIHGAKSKIIYMRT
DAAVPSIPQDMISASNSSSQLVVKWNPPSLPNGNLSYYIVRWQQQPQDRHLYQYNYCFKD
KVPNRKYANGTIDTEGGTEPTKPEGSVGEKGHYCACPKTEAEEKAEKDEAEYRKVFENFL
HNSIFVPRPNRRRRDVLAVGNSTVTSYEKNSTTEDFSNFSDSERDDIEYPFYETKVDYKW
ERTVISNLQPFTLYRIDIHSCNHEAEKLGCSASNFVFARTMPAAGADDIPGIVNTKEEDD
GVIFLGWPEPLRPNGLILMYEIEYKHQGEVHRECVSRQDYRKNGGIKLVRLPPGNYSAQV
```

```
QAISLYGNGSWTEMVSFCVKLKPDVRNNILQMVVAIPLALSFLLVGIISIVCFVFKKRNS
NRLGNGVLYASVNPEYFSAAEMYVPDKWEVPREKITMNRELGQGSFGMVYEGIAKGVVKD
EAETKVAIKTVNEAASMRERIEFLNEASVMKEFNCHHVVRLLGVVSQGQPTLVIMELMTR
GDLKSYLRSLRPDTESNSGQPTPSLKKMIQMAGEIADGMSYLNANKFVHRDLAARNCMVT
EDFTVKIGDFGMTRDIYETDYYRKGGKGLLPVRWMSPESLKDGVFTTNSDVWSFGVVLWE
IATLAEQPYQGMSNEQVLRFVMEGGLLEKPDNCPDMLFELMRMCWQFNPKMRPSFLEIIS
SIKDELDPGFKEVSFFYSEENKPPDTEELDLEAENMESIPLDPSCALQNSEHHAGHKSEN
GPGVVVLRASFDERQPYAHMNGGRKNERALPLPQSSAC

>sp|Q05688|IGF1R_BOVIN Insulin-like growth factor 1 receptor (Fragment) OS=Bos taurus GN=IGF1R PE=2
SV=1
NAIFVPRPERKRREVMQIANTTMSSRSRNTTVLDTYNITDPEELETEYPFFESRVDNKER
TVISNLRPFTLYRIDIHSCNHEAEKLGCSASNFVFARTMPAEGADDIPGPVTWEPRPENS
IFLKWPEPENPNGLILMYEIKYGSQVEDQRECVSRQEYRKYGGAKLNRLNPGNYTARIQA
TSLSGNGSWTDPVFFYVQAKTTYENFIHLMIALPIAVLLIVGGLVIMLYVFHRKRNSSRL
GNGVLYASVNPEYFSAADVYVPDEWEVAREKITMSRELGQGSFGMVYEGVAKGVVKDEPE
TRVAIKTVNEAASMRERIEFLNEASVMKEFNCHHVVRLLGVVSQGQPTLVIMELMTRGDL
KSYLRSLRPEMENNPVLAPPSLSKMIQMAGEIADGMAYLNANKFVHRDLAARNCMVAEDF
TVKIGDFGMTRDIYETDYYRKGGKGLLPVRWMSPESLKDGVFTTHSDVWSFGVVLWEIAT
LAEQPYQGLSNEQVLRFVMEGGLLDKPDNCPDMLFELMRMCWQYNPKMRPSFLEIISSVK
DEMEAGFREVSFYYSEENKPPEPEELDLEPENMESVPLDPSASSASLPLPDRHSGHKAEN
GPGPGVLVLRASFDERQPYAHMNGGRKNERALPLPQSSTC

>sp|Q29000|IGF1R_PIG Insulin-like growth factor 1 receptor (Fragments) OS=Sus scrofa GN=IGF1R PE=2
SV=2
ERTVISNLRPFTLYRIDIHSCNHEAEKLGCSASNFVFARTMPAEGADDIPGPVTWEPRPE
NSIFLKWPEPENPNGLILMYEIKYGSQVEDQRECVSRQEYRKYGGAKLNRLNPGNYTARI
QATSLSGNGSWTEPVFFYVQAKTTYENFIHLIIALPVAVLLIVGGLVIMLYVFHRKRNNS
RLGNGVMLFELMRMCWQYNPKMRPSFLEIISSIKDEMEPGFREVSFYYSEENKPPEPEEL
DLEPENMESVPLDPSASSSSLPLPDRHSGHKAENGPGPGVLVLRASFDERQPYAHMNGGR
KNER
```

2.2 ClustalW

Multiple alignments of protein sequences are important tools in studying relationships among sequences [15]. Clustal W is a general purpose multiple sequence alignment program for DNA or proteins. The alignment is progressive and considers the sequence redundancy. It produces biologically meaningful multiple sequence alignments of divergent sequences [16]. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen [17]. Default parameters are used in the analysis.

2.3 Modified Group Average Link Clustering Algorithm

A modified group average algorithm was presented in this paper.

Given a set of N items to be clustered, and an N x N distance (or similarity) matrix, the steps involved in group average clustering algorithm are:

1. *Start by assigning each item to its own cluster, so that for N items, obtain N clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain.*
2. *Calculate median distances and normalize data.*
3. *Find the closest (most similar) pair of clusters and merge them into a single cluster, to obtain cluster N-1.*
4. *Compute distances (similarities) between the new cluster and each of the old clusters. The distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster.*
5. *If the distances between two clusters are same, raise index error and go to the next average distance object.*
6. *Check the matrix is neither integer nor float. If yes, go to step 3.*
7. *Repeat steps 3-5, until all items are clustered into a single cluster of size N.*

3. RESULTS AND DISCUSSION

Hierarchical cluster analysis is a statistical method for finding relatively homogeneous clusters of cases based on measured characteristics. It starts with each case in a separate cluster and then combines the clusters sequentially, reducing the number of clusters at each step until only one cluster is left.

In this study, Euclidean distance was employed as a distance measure. Here, the distance between one cluster and another cluster is considered to be equal to the shortest distance from any member of one cluster to any member of the other cluster. A group average clustering algorithm was implemented to generate better relationships among submitted IGF-1R sequences. It compromises between Single and Complete Linkage algorithms.

The main strength of *group average* method is it is less susceptible to noise and outliers. Distances between characteristics were calculated and the nearest neighbor for each object is evaluated. Further, pairs which appear nearer to each other were segregated. The distance amid one cluster and another equaling to the average distance from any member of one cluster to the other cluster was performed. On the other hand, the distance between one cluster and another cluster was calculated to evaluate if it equals to the greatest similarity from any member of one cluster to the other, based on which clusters were generated. The output of the program reporting phylogenetic tree is given in Figure 1.

```
>>>
('RAT', 'MOUSE')
('BOVINE', 'HUMAN')
('PIG', ('BOVINE', 'HUMAN'))
(('RAT', 'MOUSE'), ('PIG', ('BOVINE', 'HUMAN')))
XENLA ------------+
                  |--
RAT --------+     |
            |--+  |
MOUSE ------+  |  |
               |--+
PIG --------+  |
            |--+
BOVINE --+  |
         |--+
HUMAN ---+
('XENLA', (('RAT', 'MOUSE'), ('PIG', ('BOVINE', 'HUMAN'))))
```

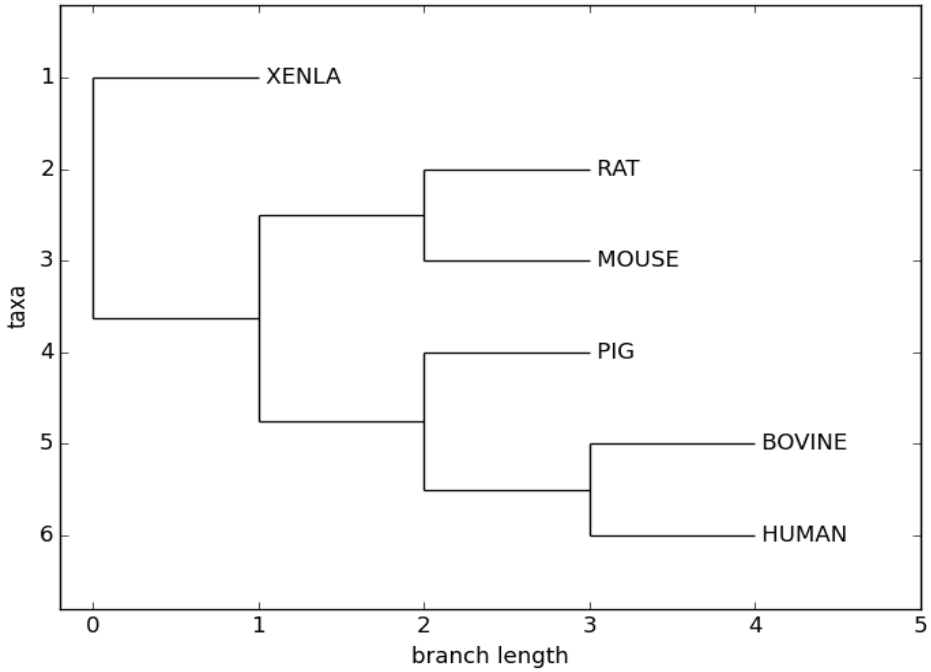

Figure 1: Phylogenetic tree obtained by group average clustering method.

From Figure 1, it can be inferred that BOVINE and HUMAN IGF1R formed one group as they both share 97.66 similarities. Comparatively, PIG IGF1R sequence shared more than 96% similarity with all sequences under study. Interestingly, it was found that all sequences shared above 90% similarity with each other except XENLA IGF1R. This organism shared around 75% similarity with all other organisms. RAT and MOUSE IGF1R sequences share 99% similarity, hence, they both appeared under one clade. IGF1R from PIG showed 96% similarity with RAT and MOUSE, hence, appeared as separate clade below.

4. CONCLUSION

Sequences can be clustered based on their similarities or dissimilarities. Hence to perform such clustering for a set of IGF1R sequences from uniprot knowledgebase, a python program was written that handles the similarity data from a set of sequences and outputs the cluster generated based on modified group average method. An example of 6 IGF1R sequences is selected for the study that resulted in clusters between all sequences. This result which is reproducible and robust was further exploited to construct phylogenetic trees based on group average clustering algorithm.

5. REFERENCES

[1] Jiang, D., Tang, C. and Zhang, A., 2004. Cluster analysis for gene expression data: a survey. *IEEE Transactions on knowledge and data engineering*, *16*(11), pp.1370-1386.

[2] https://www.cse.buffalo.edu/DBGROUP/bioinformatics/papers/survey.pdf

[3] Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M., 1999. Systematic determination of genetic network architecture. *Nature genetics*, *22*(3), pp.281-285.

[4] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, *95*(25), pp.14863-14868.

[5] Jain, A.K. and Dubes, R.C., 1988. *Algorithms for clustering data*. Prentice-Hall, Inc..

[6] Shamir, R. and Sharan, R., 2002. Algorithmic approaches to clustering gene expression data. In *In*.

[7] Jiang, T., Xu, Y. and Zhang, M.Q., 2002. *Current topics in computational molecular biology*. MIT Press.

[8] Rao, S.G. and Govardhan, A., 2014. Assessing h-and g-Indices of Scientific Papers using k-Means Clustering. *International Journal of Computer Applications*, *100*(11).

[9] Rao, S.G. and Govardhan, A., 2015. Investigation of Validity Metrics for Modified K-Means Clustering Algorithm. *i-Manager's Journal on Computer Science*, *3*(2), p.33.

[10] http://www.expasy.org

[11] Lipman, D.J. and Pearson, W.R., 1985. Rapid and sensitive protein similarity searches. *Science*, *227*(4693), pp.1435-1441.

[12] *Rossum, V.G. 2006. "*PEP 3000 -- Python 3000*". Python Software Foundation*. http://www.python.org/dev/peps/pep-3000

[13] Olson, C.F., 1995. Parallel algorithms for hierarchical clustering. *Parallel computing*, *21*(8), pp.1313-1325.

[14] Berkhin, P., 2006. A survey of clustering data mining techniques. In *Grouping multidimensional data* (pp. 25-71). Springer Berlin Heidelberg.

[15] http://www.ebi.ac.uk/clustalw

[16] Zhou, H. and Zhou, Y., 2005. SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures. *Bioinformatics*, *21*(18), pp.3615-3621.

[17] Needleman, S.B. and Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, *48*(3), pp.443-453.

6. AUTHOR PROFILE

Mr. R Rambabu is working as Associate Professor and Head of the Department, Dept. of CSE, Rajamahendri Institute of Engineering & Technology, Rajahmundry. He is pursuing Ph.D. (CSE) from JNTUK. He received his M.Tech (IT) from Andhra University. His areas of interest are Bioinformatics, Data Mining, Computer Networks, Software Engineering, and Mobile Computing technologies.

Dr. Peri Sriniasa Rao is a Professor at the Department of Computer Science & Systems Engineering, Andhra University Visakhapatnam. His research interest is in the areas of Image Processing, Queuing Applications, Bioinformatics and Computer Networks. At AU additionally, he is also a Chairman Board of Studies and Expert Member in CSE for AICTE and NBA. He guided 15 students for Ph.D. Degree. He is a Life Time member of CSI.