DATA REDUCTION TECHNIQUES FOR HIGH DIMENSIONAL BIOLOGICAL DATA

Shyam Mohan J S[1], P.Shanmugapriya[2], N.Kumaran[3]

[1]*Research Scholar, Department of CSE, SCSVMV University, Enathur, Kanchipuram - 631 561*
[2]*Associate Professor, Department of IT, SCSVMV University, Enathur, Kanchipuram - 631 561*
[3]*Assistant Professor, Department of IT, SCSVMV University, Enathur, Kanchipuram - 631 561*

Abstract

High dimensional biological datasets in recent years has been growing rapidly. Extracting the knowledge and analyzing high-dimensional biological data is one the key challenges in which variety and veracity are the two distinct characteristics. The question that arises now is, how to perform dimensionality reduction for this heterogeneous data and how to develop a high performance platform to efficiently analyze high dimensional biological data and how to find the useful things from this data. To deeply discuss this issue, this paper begins with a brief introduction to data analytics available for biological data, followed by the discussions of big data analytics and then a survey on various data reduction methods for biological data. We propose a dense clustering algorithm for standard high dimensional biological data.

Keywords: *Big Data Analytics, Dimensionality Reduction.*

----------***----------

1. INTRODUCTION

Data Reduction is a part of the analysis where the data is shortened, focused and organizing in such a way that Researchers take effective decisions based on the data pulled out. Dataset reduction can be done in many ways .First, data coded in an exploratory analysis. Second, data partitioning for theoretical or hypothesis testing. Third, by eliminating the irrelevant data which is not used for analysis.[1]. High Dimensional data sets in biology arise in many ways. With the widespread utilization of DNA microarrays and high throughput, parallel analysis of thousands of genes for gene expression data has been done easily. [2].One of the common technique is feature extraction and dimensionality reduction.[3].Datasets collaborate with each other (also called Big data) have a large number of dimensions (attributes) .[4, 5].Each dataset represents a point in a multidimensional space. Clustering is an unsupervised learning method which do not require any supervision, i.e., it groups similar data points into clusters (dense regions).[6]. As the dimensionality of data increases, it is difficult to form clusters. Traditional clustering algorithms like DBSCAN [7] generate clusters in the full-dimensional space by measuring the proximity and dimensions of a dataset and therefore these classical clustering algorithms are ineffective as well as inefficient for the high-dimensional datasets. [8]. In data mining, data reduction for low dimension datasets can be done using principal component analysis [9], singular value decomposition[10], and dynamic tensor analysis[11].but however the above techniques fail for high dimension datasets. Generally, high-dimensional data [12] has two main implications: (1) Finding the relative contrast between similar and dissimilar points is difficult as the dimensionality of data grows.[13]. (2) Grouping of different data sets with each other.[14].

2. MOTIVATING EXAMPLES

High dimensional Data sets in biology: In biology, high dimensional data sets are obtained from microarray chips, which forms a matrix [15].Each cell in the matrix contains rows and columns , rows contain the expression level of a gene and columns contain experimental condition. Rows and columns can be clustered to form meaningful biological inferences. But, only a subset of genes are considered for experiments.[16].Forming clusters for biological data is called as bi clustering for microarray datasets.[17].

Bioinformatics: In bioinformatics, large genomes are represented using Expressed Sequence Tag (EST) that contain portions of genes (mature mRNA) of length at least 500–800 nucleotides long. Finding locations of genes can be performed by searching.[18]. The task of identifying related or different genes or protein domains by sequence search techniques is a complex task. A highly beneficial search for proteins is Remote homology.[19,20].

3. BACKGROUND AND LITERATURE WORK

Many data reduction techniques have been proposed in the literature. Data reduction techniques can be divided into two categories: global alignment and local alignment methods. Global alignment or full-sequence matching is used in global optimization, and local alignment method is used in identifying the local regions that are globally divergent. Table 1 shows various alignment methods available with their description.

Table 1: Existing Data Reduction Techniques

SI. No.	Name	References	Description
1.	BLAST	21	Uses heuristic search for computing long biological sequences.
2.	BLAST2	22	Performs limited number of insertions and deletions to increase the speed of computation.
3.	MegaBLAST	23	Uses greedy search algorithm for detecting a variety of sequences.
4.	MPBLAST and miBLAST	24	Provide results for queries done in parallel.
5.	BLAT	25	Provides indexing and linear search for finding the best match in the index.
6.	OASIS	26	Performs similarity measure using local alignment technique.
7.	The Shift-Or algorithm	27	Used in searching large queries by following bit manipulation technique.
8.	RBSA	28	Uses alphabetic reduction technique in which number 1 represents odd characters and 0 represents even letters.
9.	short-read sequencing methods	29	Performs near-exact subsequence matching comparison techniques by assumption. Ex: SOAP
10.	q-gram	30	Follows pigeon-hole principle i.e., two sequences share similar or distinguishing sequences .

For parallel processing analytics, Principal Component Analysis (PCA) is a well known model for dimensionality reduction which requires few number of I/O and CPU operations.[31,32, 33]. It can be applicable to any data set with numeric dimensions. It is used in data mining for dimensionality reduction.

EntityRel [34], is a entity correlation graph(biomedical graph) constructed from unstructured data by computing the differences between two heterogeneous entities .It extends the scope of unstructured text data by finding the meta path based relationship analysis [35] using data mining techniques.

DRESS, a dimensionality reduction technique used in finding efficient sequence search is capable of handling large queries and provides accurate results in optimal time. Applies mapping transformations for original strings toa new strings for dimensionality reduction.[36]

IHOSVD is a tensor based dimensionality reduction method which applies recursively the incremental matrix decomposition algorithm and periodically updating the orthogonal bases for computing the new or the core tensor.[37].In tensor based dimensionality reduction method, data is categorized into various modules like data: collection, tensorization, dimensionality reduction, analysis and service modules. We apply clustering techniques to one of the module in the above model.

4. RELATED WORK

4.1 Data Tensorization Module

In this module, the collected data is not uniform i.e, it is usually unstructured, semi-structured or structured data .In a unified tensor model, the above data is represented in various sub-tensors and finally combined to form a unified heterogeneous tensor. Most of the unstructured data includes video and audio data. Similarly XML documents, ontological data, etc. are semi-structured data and structured data is a combination of numbers and strings.[37].For biological data, we consider semi-structured data in a subspace cluster.

Definition: Clustering in Subspace

Detecting clusters in free space or subspace is clustering in subspace. Points may be members belonging to any multiple clusters existing in free space. [38].

Problem statement: For every 2^d different subspaces in subspace clustering, it is given by the fact that there is a space with d dimensions.

Preliminary

For any point p, we define core or center points, nearby points and outliers that are clustered together. For any point p there exists three points as given below:

- Point p is a center point if any of the minimum points are within distance x and which are nearby to p. No points are nearby from a non-center point.
- A point q is nearby to p if there exists a path p_1, ..., p_n where p_1 = A and p_n = B, such that for each p_{i+1} p is nearby to p_i i.e., all the points on the path must be center points excluding B.
- Far away points which could not be reached are called outliers.

As discussed in the above statement, any point p forms a cluster consisting of center points joining other non-center points that are nearby to the center point. Hence each cluster consists of at least one center point.

XML Example 1:[39]

<Gene><genes><gene id="361"><name>BNGT1</name><organism>Human</organism><annotations><protein_sequence>

BEPCKMHCNMDHNKG

</protein_sequence></annotations></gene><gene

id="362"><name>BNGT2</name><organism>Human</organism><family>angiopoietin</family><annotations><protein_sequence>

KQWJIHFTLHCILM
</protein_sequence></annotations></gene></genes><interactions><sequence_similarity id="10001" gene1="361" gene2="362"><score> 70
</score></sequence_similarity></interactions><relationships/></Gene>

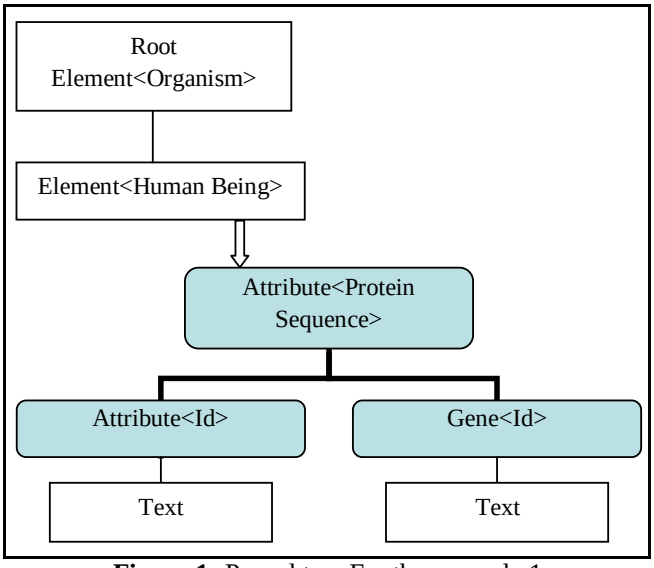

Figure 1: Parsed tree For the example 1

5. METHOD

Without loss of generosity, we assume that for each individual cluster we have one data point in step 1. In step 2, the two neighboring clusters are merged to form a new cluster leaving one cluster behind and therefore forming a hierarchical clustering structure. STREAM algorithm is used for constructing hierarchical clustering structure.[40] In STREAM, the original data is divided [41]into clusters and then the small sized clusters are merged with bigger clusters using agglomerative algorithm.

5.1 Clustering Algorithm

Input: Initialize cluster size say D=30, distance x , minimum points
Output: All P' x-nearby points and P.

1. Assign C to 0
2. Begin
3. Move For each point in dataset D
4. If(P=visited)
5. Proceed to next point which is nearby
6. mark P =visited
7. Nearbypoints1 = Compute Query(P, x)
8. If((size of NeighborPts1) < minimum points)
9. Then
10. Mark P =Outlier
11. else
12. Move C to next cluster

Expanding Cluster
Input: Initialize cluster size say D=30, distance x , minimum points
Output: All P' x-nearby points and P.

1. Read Cluster C, Nearby points, P, x, minimum points
2. Move P to cluster C
3. For each point P' in Nearby points1 and
4. If P' is not visited Then
5. Mark P'=visited
6. Nearby points2 = Compute Query(P', x)
7. If((size of Nearby points2) >= minimum points) Then
8. Nearby points = Nearby points1 joined with Nearby points2
9. If P ∄Cluster C
10. Add P' to cluster C
11. Computed values(P, x)

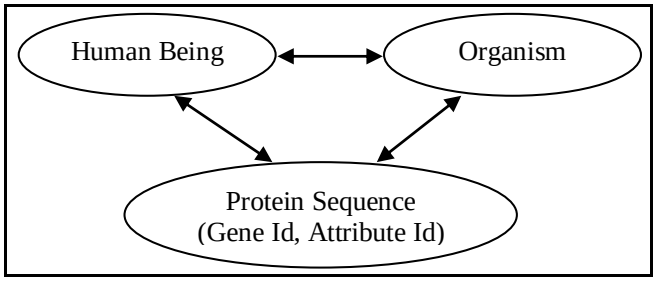

Figure 2: Clusters for the sample dataset

Table 2: Keywords and their Description

Keywords	Description
ProbeID	Unique identifier for each probe sequence
Annotation	miRNAs recognized by the probe
miRNA h	Human
miRNA m	Mouse
miRNA r	Rat
Sequence	sequence of probe

6. RESULTS

6.1 Datasets

In this paper, we apply clustering techniques for dimensionality reduction for biological data. The miRNA sequences can be downloaded from [42].The statistical results for 30 samples of miRNA sequences is shown in table 3. For evaluation, we took 30 sample datasets from table 3.The Iris dataset[43] which can be downloaded from the UCI machine learning repository[44] and the Rat Central Nervous System (CNS) dataset can be downloaded from [45].The Iris dataset contains 50 samples with four features measured from each sample .[46].The Rat CNS dataset consists of 112 genes over 17 conditions during rat central nervous system development.[47].We apply clustering algorithm for miRNA data sets. Each datasets is provided with a probe Id followed by annotation and the sequence. For other datasets we just display the final number of clusters in table 4.The final result is three clusters with their sample datasets shown in figure.2.

6.2 Standard Biological Datasets

We applied our clustering algorithm for high dimensional biological data. Each dataset consists [48]of 500 samples(apprx.) stored over each dataset. For evaluation, we take the datasets in XML format which is also used in tensor based dimensionality reduction technique.

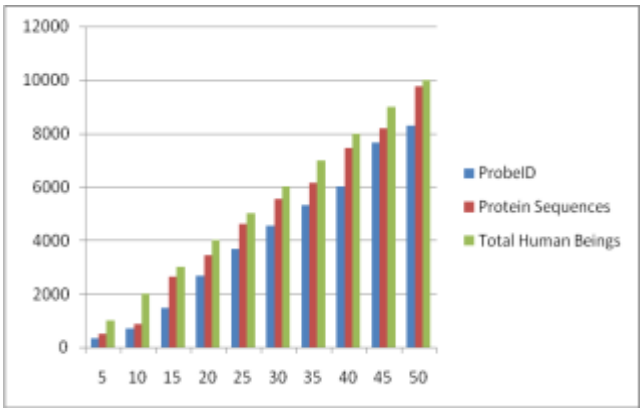

Figure 3: Rapid growth of miRNA Sequences

Table 4: Total number of Clusters for Standard High dimensional Biological Datasets.		
Dataset	Number of Samples Taken	Number of Clusters
Iris dataset	50	2
miRNA Dataset	130(apprx.)	3
Rat CNS dataset	112 genes over 17 conditions	6
cell phenotype dataset	444 samples in 41dimensions with 4 centers	5

7. CONCLUSION

Forming Clusters for large biological datasets is a difficult task which involves many algorithms. Real time video streaming of biological datasets cannot be computed easily as they contain highly sensitive information. In future we hope to compute real time video analytics for biological datasets.

Table 3: Statistical Results of miRNA Sequences		
Protein Coding Sequence ID	Annotation	Protein Sequence of max. length 30
EamA190	h-miR-10b_rfam7.0	PKKMLECCTRCGEYIAMHSIYWQVLDWHAF
EamA187	hmr-miR-107_rfam7.0	AKKMLCKCTRCMEYIAMKSIYWQVLDWHE
EamA185	hmr-miR-103_rfam7.0	AKOMLECCHRCGEYIAMHSIYWQVLDWHA
EamA181	hmr-let-7f_rfam7.0	PBKMLECCTRIGEYIXMHSIYWZVLDWHAF
EamA179	hmr-let-7d_rfam7.0	MAMLECCTRCGEYIAMHSIYWQVLDWHAG
EamA177	mr-miR-101b_rfam7.0	MAHMLECCTRCGEYIAMHSIYWQVLDWHA
EamA175	hmr-miR-320_rfam7.0	KHGMLECCTRCGEYIAMHSIYWQVLDWHAF
EamA168	hmr-let-7e_rfam7.0	PKKMLECCTRFTHYIAMHSIYWQVLDWHAF
EamA161	hmr-miR-28_rfam7.0	SHYTLECCTRCGEYIAMHSIYWQVLDWHAF
EamA160	hmr-miR-26b_rfam7.0	HJYTREYUTRCGEYIAMHSIYWQVLDWHAF
EamA155	hmr-miR-136_rfam7.0	AKKMLCKCTRCMEYIAMKSIYWQVLDWHEF
EamA283	mr-miR-211_rfam7.0	AKOMLECCHRCGEYIAMHSIYWQVLDWHAF
EamA282	m-miR-199b_rfam7.0	PBKMLECCTRIGEYIXMHSIYWZVLDWHAFE
EamA281	mr-miR-217_rfam7.0	MAMLECCTRCGEYIAMHSIYWQVLDWHAG
EamA280	hmr-miR-30a-3p_rfam7.0	MAHMLECCTRCGEYIAMHSIYWQVLDWHAK
EamA279	hmr-miR-29c_rfam7.0	KHGMLECCTRCGEYIAMHSIYWQVLDWHAF
EamA278	hmr-miR-98_rfam7.0	PKKMLECCTRFTHYIAMHSIYWQVLDWHAF
EamA238	hm-miR-1_rfam7.0	SHYTLECCTRCGEYIAMHSIYWQVLDWHAF
EamA270	hmr-miR-30b_rfam7.0	HJYTREYUTRCGEYIAMHSIYWQVLDWHAF
EamA159	hmr-miR-130a_rfam7.0	MAHMLECCTRCGEYIAMHSIYWQVLDWHAK
EamA163	hmr-miR-142-3p_rfam7.0	KHGMLECCTRCGEYIAMHSIYWQVLDWHAF
EamA171	hmr-miR-137_rfam7.0	PKKMLECCTRFTHYIAMHSIYWQVLDWHAF
EamA306	m-miR-201_rfam7.0	SHYTLECCTRCGEYIAMHSIYWQVLDWHAF
EamA307	m-miR-202_rfam7.0	HJYTREYUTRCGEYIAMHSIYWQVLDWHAF
EamA308	hmr-miR-206_rfam7.0	AKKMLCKCTRCMEYIAMKSIYWQVLDWHEF
EamA309	m-miR-207_rfam7.0	AKOMLECCHRCGEYIAMHSIYWQVLDWHAF
EamA310	hmr-miR-208_rfam7.0	PBKMLECCTRIGEYIXMHSIYWZVLDWHAFE
EamA247	hmr-miR-212_rfam7.0	MAMLECCTRCGEYIAMHSIYWQVLDWHAG
EamA251	hmr-miR-216_rfam7.0	MAHMLECCTRCGEYIAMHSIYWQVLDWHAK
EamA253	hmr-miR-218_rfam7.0	KHGMLECCTRCGEYIAMHSIYWQVLDWHAF

REFERENCES

[1].“Handbook For Team-Based Qualitative Research”, Greg Guest and Kathleen M. MacQueen, Altamira Press, United Kingdom, 2008.

[2]. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270(5235):467–470.

[3]. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotech.26(10):1135–1145.

[4]. Fan J, Han F, Liu H (2014) Challenges of big data analysis. National Science Review 1(2):293–314.

[5]. Steinbach M, Ertöz L, Kumar V (2004) The challenges of clustering high dimensional data. In: New directions in statistical physics. Springer, Berlin Heidelberg. pp 273–309.

[6].Aggarwal CC, Reddy CK (2013) Data clustering: algorithms and applications. Data Mining Knowledge and Discovery Series 1st. CRC Press.

[7]. Ester M, Kriegel H, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. Int Conf Knowl Discov Data Min 96(34):226–231.

[8]. Zhang T, Ramakrishnan R, Livny M (1996) BIRCH: an efficient data clustering method for very large databases. In: Proc. of the ACM SIGMOD international conference on management of data, vol. 1. ACM Press, USA. pp 103–114.

[9].H. Abdi and L. J. Williams,``Principal component analysis," Wiley Interdiscipl. Rev., Comput. Statist., vol. 2, no. 4, pp. 433_459,Jul./Aug. 2010.

[10]. M. Brand, ``Incremental singular value decomposition of uncertain data with missing values," in Proc. 7th Eur. Conf. Comput. Vis. (ECCV), 2002, pp. 707_720.

[11]. J. Sun, D. Tao, and C. Faloutsos, ``Beyond streams and graphs: Dynamic tensor analysis," in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), 2006, pp. 374_383.

[12]. Bellman RE (1961) Adaptive control processes: a guided tour. Princeton University Press, New Jersey

[13]. Beyer K, Goldstein J (1999) When is nearest neighbor meaningful? Proc 7th Int Conf Database theory. In: DatabaseTheory –ICDT’99. Lecture Notes in Computer Science. Springer, Berlin Heidelberg Vol. 1540. pp 217-235.

[14]. Parsons L, Haque E, Liu H (2004) Subspace clustering for high dimensional data: a review. ACM SIGKDD Explor News l6(1):90–105.

[15]. Babu MM (2004) Introduction to microarray data analysis. In: Grant RP (ed). Computational genomics: Theory and application. Horizon Press, UK. pp 225–249

[16]. Jiang D, Tang C, Zhang A (2004) Cluster analysis for gene expression data: a survey. IEEE Trans Knowl Data Eng 16(11):1370–1386

[17].Cheng Y, Church GM (2000) Biclustering of expression data. Proc Int Conf Intell Syst Mol Biol 8:93–103

[18].Jongeneel CV (2000) Searching the expressed sequence tag (est) databases: panning for genes. Bioinformatics 1:76–92.

[19].Liu B, Wang X, Zou Q, Dong Q, Chen Q (2013) Protein remote homology detection by combining chous pseudo amino acid composition and profile-based protein representation. Mol Inf 32(9–10):775–782.

[20]. Bhadra R, Sandhya S, Abhinandan KR, Chakrabarti S, Sowdhamini R, Srinivasan N (2006) Cascade psiblast web server: a remote homology search tool for relating protein domains. Nucleic Acids Res 34(Web–Server–Issue):143–146.

[21].Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410.

[22]. Altschul S, Madden T, Schffer R, Zhang J, Zhang Z, MillerW, Lipman D (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

[23]. Zhang Z, Schwartz S,Wagner L,MillerW(2000) A greedy algorithm for aligning dna sequences. J Comput Biol 7:203–214.

[24].Korf I, Gish W (2000) Mpblast : improved blast performance with multiplexed queries. Bioinformatics 16:1052–1053.

[25].Kent WJ (2002) Resource BLAT-The BLAST-like alignment tool. Genome Res.

[26].Meek C, Patel JM, Kasetty S (2003) Oasis: an online and accurate technique for local-alignment searches on biological sequences. In: Proceedings of very large database endowment (PVLDB), vol 29, pp.910–921.

[27]. Baeza-Yates R, Gonnet GH (1992) A new approach to text searching. Commun ACM 35(10):74–82.

[28]. Papapetrou P, Athitsos V, Kollios G, Gunopulos D (2009) Reference-based alignment in large sequence databases. Proc Very Large Database Endow (PVLDB) 2(1):205–216.

[29]. Li R, Li Y, Kristiansen K, Wang J (2008c) Soap: short oligonucleotide alignment program. Bioinformatics 24(5):713–714.

[30].Yang X, Wang B, Li C (2008) Cost-based variable-length-gram selection for string collections to support approximate queries efficiently. In: Proceedings of the 2008 ACMSIGMOD international conference on Management of data, ACM, pp 353–364.

[31].Carlos Ordonez Et.Al.” PCA for large data sets with parallel data summarization” Distrib Parallel Databases (2014) 32:377–403,Springer, DOI 10.1007/s10619-013-7134-6.

[32].Halko, N.H., Martinsson, P.-G., Shkolnisky, Y., Tygert, M.: An algorithm for the principal component analysis of large data sets. SIAM J. Sci. Comput. 33(5), 2580–2594 (2011).

[33].Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning, 1st edn. Springer, New York (2001).

[34].Ming Ji Et.al,” Mining strong relevance between heterogeneous entities from unstructured biomedical data”, Data Min Knowl Disc (2015) ,Springer,29:976–998. DOI 10.1007/s10618-014-0396-4.

[35].SunY,Han J,YanX,Yu PS,Wu T (2011) Pathsim: meta path-based top-k similarity search in heterogeneous information networks. PVLDB 4(11):992–1003.

[36].Alexios Kotsifakos et.al.” DRESS: dimensionality reduction for efficient sequence search”, Data Min Knowl

Disc (2015) 29:1280–1311,Springer, DOI 10.1007/s10618-015-0413-2.
[37]. LIWEI KUANG Et.Al." A Tensor-Based Approach for Big Data Representation and Dimensionality Reduction", IEEE Transactions on emerging Topics in Computing, Volume 2,No.3, September 2014.
[38]. https://en.wikipedia.org/wiki/DBSCAN.
[39].http://authors.phptr.com/graves/book_examples/ch10/ex10_2.xml.
[40].O'callaghan L, Meyerson A, Motwani R, Mishra N, Guha S (2002) Streaming-data algorithms for high-quality clustering. In: ICDE, p 0685
[41]. Karypis G, Han E, Kumar V (1999) Chameleon: hierarchical clustering using dynamic modeling. Computer 32:68–75.
[42].http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi.
[43].Fisher, R. A. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179–188 (1936).
[44].Bache,K. & Lichman, M. Uci machine learning repository, 2013. URL http://archive. ics. uci. edu/ml (1990).
[45].Wen, X. et al. Large-scale temporal gene expression mapping of central nervous system development. Proceedings of the National Academy of Sciences 95, 334–339 (1998).
[46].Chenyue W. Hu Et.Al.," Progeny Clustering: A Method to Identify Biological Phenotypes", Open Scientific Reports, DOI: 10.1038/srep12894.
[47]. Giancarlo, R., Scaturro, D. & Utro, F. Computational cluster validation for microarray data analysis: experimental assessment of clest, consensus clustering, figure of merit, gap statistics and model explorer. BMC Bioinformatics 9, 462 (2008).
[48].Linxia Wan Et. Al." Automatically clustering large-scale miRNA sequences: methods and experiments", International Conference on Intelligent Biology and Medicine (ICIBM) Nashville, TN, USA. 22-24 April 2012, http://www.biomedcentral.com/1471-2164/13/S8/S15.