Survey : Clustering techniques for large-scale genomic data

Sarthak Satpathy
sarthak.satpathy@emory.edu
Emory University
Atlanta, GA, USA

ABSTRACT

The survey discusses the different clustering algorithms used to derive inferences from genomic data. Through decades of development in the field of unsupervised learning, there has been an evolution in the techniques. However, the use of traditional approaches on these high dimensional, extremely sparse data might not be beneficial. The focus here would be to discuss the commonly used algorithms, their limitations and performances in the context of sequencing experiments.

KEYWORDS

clustering, sequencing, genomics, graph-based

ACM Reference Format:
Sarthak Satpathy. 2018. Survey : Clustering techniques for large-scale genomic data. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Clustering is an unsupervised learning technique that aims to group data into partitions where the members in the same cluster have a higher similarity. In contrast, members in different clusters have higher differences. [19] Clustering process can be divided into four primary steps: feature extraction and selection, clustering algorithm design, evaluation and explanation. [**?**]

It has been observed that the amount of biological data doubles every 18 month. For example the amount of sequences in GenBank's nucleic acid database has evolved from 606 sequences in 1986 to 162 million sequences in 2013. Large scale sequencing data range from transcriptomics, single-cell transcriptomics, spatial transcriptomics and genomics experiments. The data tend to be high dimensional. With an increase in the resolution and complexity of the sequencing technique, the number of features exponentially increases. Clustering tends to be the primary tool for exploring this data to observe patterns at different scales of biological inferences: gene, protein , cell, sample, species, etc.

Due to the sparse, high dimensional nature of sequencing data, using traditional clustering algorithms provide poor performance when compared to recent graph-based clustering approaches. While there are surveys that aim at comparing different clustering algorithms through the decades, it is important to note that the rapid diversification of the sequencing techniques demand more closer look at the algorithms that can be used on these data. [7, 8, 12] This survey would highlight the performances and limitations of the algorithms, their modifications and usage in inferring biological phenomenon from sequencing data. The evaluation and selection criteria for the techniques would also be discussed in the following sections.

Even the most basic application of unsupervised learning in clustering genes from gene-expression data has seen the change in sequencing from probe-based microarray experiments to current single-cell and spatial transcriptomics experiments. While it was important to know the gene clusters before, now the techniques require to cluster based on cell and gene level. This increased scale with an increase in sparsity does challenge the existing techniques. The traditional methods of clustering included heirarchical method, k-means based , model-based methods, soft clustering, grid based and density based techniques. Recently, graph-based approaches like k-clique communities, maximal clique, CAST, CLICK, etc have gain popularity. [7]. Neural network based methods like self organizing maps are example of other approaches, recently used in the field.

In the survey we would discuss the different sequecing data and why they need clustering algorithms to give biological insights to the problem. The different techniques would be elaborately discussed where they would be compared on the metrics of memory requirement, storage, time complexity and technical limitation with respect to the experiments.

The survey can serve as a manual for people working in the field to choose the algorithm in a context-specific manner. It will also bring out the current challenges in the field that needs redressal in terms of modifications to earlier algorithms.

2 CHALLENGES PRESENTED BY BIOLOGICAL DATA

Before jumping to the clustering algorithms, it would be convenient to discuss the challenges that biological data presents compared to other forms of data. Large datasets make it easier for accurate prediction from machine learning algorithms. Currently, the magnitude of most biological data is too small. Although the number of features per single experiment or sample is large, the number of samples or replicates is still tiny for a more straightforward model prediction. This situation is commonly called the "small n and large p" problem. [16] The total amount of biological data is enormous and is dynamically increasing. However, it is most often sourced from different platforms. The differences in technology and experimental conditions make it challenging to move towards an integrative approach. [21]

Due to the heterogeneity in biological data, models trained over specific data may not be well generalized for other data. If the

CS-570, January 23, 2022, Emory University

https://doi.org/10.1145/1122445.1122456

validation data is significantly different from the training data, the analysis results of the clustering algorithm may not be interpretable.

The clustering algorithms have a black-box nature in model development. [21] From a biological point of view, it is usually difficult to interpret or find mechanisms from the model output. The end-users of the applications of these algorithms tend to be wet-lab scientists who have little or no knowledge of the details of the process. The demand in such scenarios is to understand the biological or physiological basis of the separation of datasets to make real-world interpretations.

Working with genomic data requires handling many candidate targets that are often crowded by false positives(FP). A 5% p-value threshold on any statistical test would mean in a set of ten thousand genes; we would most likely observe 500 FP genes which the algorithm can misclassify. A strategy to make the selection criteria conservative, for example, to a 1% or lower threshold, would lead to many false negatives. This is commonly referred to as the multiple comparison issue. Although there are traditional methods of the p-value adjustments, they often end up being very conservative due to a large number of tests. A balance between the Type I and Type II errors can help solve the problem. [10, 16, 18]

High dimensional nature of data in the biological context leads to sparsity in the data space. Mathematical and computational approaches have limitations in capturing high dimensional data efficiently. Dimension reduction is the most common strategy here. However, users need to be careful that observations from several lower dimensions might not be simlar to the phenomenon explained by the joint, high-dimensional data. [6, 10]

Computational limitation often comes up while solving genomic problems. Many problems in genomic data analysis have been proven to be theoretically of non-polynomial)-hard complexity which means that no computational algorithm can search for all possible candidate solutions. Hence, the development of heuristic approaches based on appropriate probabilistic modeling and statistical inference is the way forward to deal with the computational limitation of genomic data. [10, 15]

Noise in the data is an inherent characteristic of high-throughput biological data. Most often the signal of interest is confounded by other factors in the experiment. The experimental design can introduce error and bias. Hencem, separating the biological noise from technical noise remains a challenge in such technologies. There is no definitive strategy to deal with this problem. Most oftem people rely on domain expertise and existing distributions while processing such noisy data. [10]

3 DISTANCE AND SIMILARITY MEASURES

A clustering approach is first expounded by a similarity or dissimilarity measure or distance index. [10] Distance functions are commonly used for quantitative data features while similarity is preferred for qualitative data features. [20]

Distance functions:

3.1 Minkowski distance

Minkowski function is a generralised function for the power n. [Equation 1] Based on the value of n following are the definitions for distance, City-block distance when n = 1, Euclidean distance when n = 2 and Chebyshev distance when n $\rightarrow \infty$

$$\left(\sum_{l=1}^{d} |x_{il} - x_{jl}|^n\right)^{1/n} \tag{1}$$

3.2 Standardized Euclidean distance

A weighted Euclidean distance based on the deviation.

$$\left(\sum_{l=1}^{d} \left|\frac{x_{il} - x_{jl}}{s_l}\right|^2\right)^{1/2} \tag{2}$$

where s stands for the standard deviation.

3.3 Cosine distance

Cosine distance remains same in face of the rotation change of data.

$$1 - \cos\alpha = \frac{x_i^T x_j}{\|x_i\| \left\|x_j\right\|} \tag{3}$$

3.4 Pearson correlation distance

Pearson correlation distance measures the distance based on linear correlation.

$$1 - \frac{Cov\left(x_i, x_j\right)}{\sqrt{D\left(x_i\right)}\sqrt{D\left(x_j\right)}} \tag{4}$$

where Cov stands for the covariance for and D stands for the variance.

3.5 Mahalanobis distance

It is commonly used with high computation complexity.

$$\sqrt{\left(x_i - x_j\right)^T S^{-1} \left(x_i - x_j\right)} \tag{5}$$

Similarity functions

3.6 Jaccard similarity

Measure the similarity of two sets.

$$J\left(A, B\right) = \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

where |X| stands for the number of elements of set X.
$Jaccarddistance = 1 - Jaccardsimilarity$

3.7 Hamming similarity

The minimum number of substitutions needed to change one data point into the other. The number is smaller, the similarity is more. It is most commonly used for string data.

3.8 For data of mixed type

Strategies include to map the features into (0,1). Transform the features into dichotomous ones. [5]

$$S_{ij} = \frac{1}{d}\sum_{l=1}^{d} S_{ijl} \tag{7}$$

$$S_{ij} = \left(\sum_{l=1}^{d} \eta_{ijl} S_{ijl}\right) / \left(\sum_{l=1}^{d} \eta_{ijl}\right) \tag{8}$$

4 METHODS FOR MICROARRAY DATA

The past decades has seen a revolution in sequencing technologies. The transition from microarray to Ribonucleic Acid Sequencing using Next-Generation Sequencing technologies has not only made the investigation of genomic data easier but also cheaper. Many serveys in the past decades have focussed on microarray based genomic data. In this section we would brush through the commonly used methods which were later modsified or taken up for the RNA-seq data. In the early 2000s, microarray technologies had made it possible to monitor the expression levels for tens of thousands of genes in parallel. A microarray experiment would involve the binding of DNA sequence to a known probe. The probe would be fluorescently or radioactively labelled. The chip on which the DNA and probe is scanned and the intensity of light for each grid gives the expression for the gene. A gene expression data set from a microarray experiment can be represented by a real-valued expression matrix. Figure 1 is a representation of the expression matrix. The w_{ij} cell represents the gene expression for gene i in sample j.

Two classes of clustering algorithms have been traditionally used in genomic data analysis: hierarchical and partitioning allocation algorithms. Hierarchical algorithms that allocate each subject to its nearest subject or group. Partitioning algorithms that divide the data into a pre-specified number of subsets. [10]

4.1 K-Means

K-means is a partition algorithm. [17] Given a prespecified number K, the algorithm partitions the data set into K disjoint subsets which optimize the following objective function:

$$E = \sum_{i=1}^{K} \sum_{O \in C_i} |O - \mu_i|^2 \tag{9}$$

where, O is a data object in cluster C_i and μ_i is the centroid (mean of objects) of C_i. Thus, the objective function E tries to minimize the sum of the squared distances of objects from their cluster centers.

4.1.1 Advantages. : Algorithm is simple and fast. The algorithm converges in a small number of iterations.

4.1.2 Disadvantages. : Number of gene clusters is usually unknown in advance. Fine tuning of parameters can be an exhaustive process due to large number of combinatorials. K-means algorithm is sensitive to the noise in the gene expression data.

4.2 Self-Organizing Map

The Self-Organizing Map (SOM) was created using a single layered neural network. Each neural network neuron has a reference vector associated with it, and each data point is "mapped" to the neuron with the "closest" reference vector. All data points are mapped to the output neurons to identify clusters.

4.2.1 Advantages. : Generates an appealing map of high-dimensional data set in 2D or 3D space and places similar clusters near each other. The method is robust in dealing with noise compared to k-means.

4.2.2 Disadvantages. : The parameters supplied to the algorithm are used throughout. Hence, improper specification can lead to non-discovery of natural clusters.

Hierarchical Clustering

Hierarchical clustering generates a hierarchical series of nested clusters which can be visually represented by a tree, called dendrogram. There are two types of hierarchical clustering approaches: agglomerative approaches (bottom-up) and divisive approaches(top-down).

4.3 UPGMA

UPGMA (Unweighted Pair Group Method with Arithmetic Mean) is a method from Eisen et al. [2] Fluorescence ratio decides the color of each cell of the matrix. The rows are reordered based on the dendrogram structure and a consistent node-ordering rule. At the end, color patches that are large and contiguous signify groupings of genes that have comparable expression patterns.

4.4 DAA

Deterministic-annealing algorithm (DAA) [1]. The expression pattern of gene k was represented by a vector $\vec{g}_k$, and the probability of gene k belonging to cluster j was assigned according to a two-component Gaussian model:

$$P_j(\vec{g}_k) = exp(-\beta|\vec{g}_k - C_j|^2) / \sum_j exp(-\beta|\vec{g}_k - C_j|^2) \tag{10}$$

The cluster centroids were recalculated by.:

$$C_j = \sum_k \vec{g}_k P_j(\vec{g}_k) / \sum_k P_j(\vec{g}_k) \tag{11}$$

Advantages: Groups together genes with similar expression pattern and provides a graphical way to represent data.

Disadvantages. : Small perturbation in data can change the structure of dendogram. Computational complexity and greedy nature of the algorithm are other disadvantages,

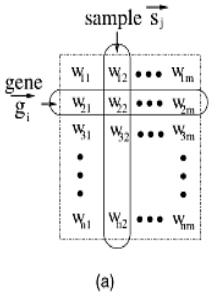
4.5 Graph-Theoretical Approaches

CLICK (CLuster Identification via Connectivity Kernels) and CAST were the commonly used graph based algorithm for dealing with microarray data.

In future surveys on microarray data where the runtime and Jaccard similarity was used to compare clustering efficiencies it was observed that, for small clusters, maximal clique and paraclique work best; for medium clusters, k-clique communities and paraclique work best. For big clusters, Ward and CAST are the most effective approaches.

5 METHODS FOR SINGLE-CELL RNA-SEQ

RNA-Seq expression matrices are very similar to microarray and the algorithms can be directly translated for clustering purposes. The single-cell data is a highle resolved, high dimension and largely



n	number of genes
m	number of samples
M	a gene expression matrix
w_{ij}	each cell in a gene expression matrix
$\vec{g_i}$	a gene
$\vec{s_j}$	a sample
$G, G', G_0, ...$	a set of genes
$S, S', S_0, ...$	a set of samples

(b)

Figure 1: Microarray data representation Jiang et al. [8]

sparse expression data which needs a lot of fine tuning for clustering purposes. Here are a few algorithms used currently and their approach. [3]

5.1 ascend (v0.5.0)

PCA dimension reduction (dim=30) and iterative hierarchical clustering. [14]

5.2 CIDR (v0.1.5)

PCA dimension reduction based on zero-imputed similarities, followed by hierarchical clustering. [11]

5.3 FlowSOM (v1.12.0)

PCA dimension reduction (dim=30) followed by self-organizing maps and hierarchical consensus meta-clustering to merge clusters. [4]

5.4 SC3svm (v1.8.0)

PCA dimension reduction or Laplacian graph. K-means clustering on different dimensions. Hierarchical clustering on consensus matrix obtained by K-means. Support vector machine (SVM) to classify the rest. [9]

5.5 Seurat (v2.3.1)

Dimension reduction by PCA (dim=30) followed by nearest neighbor graph clustering. [13]

6 CONCLUSIONS AND FUTURE DIRECTIONS

The aim of this survey was to introduce experimental scientists to the advantages and disadvantages of various clustering algorithms. We discussed the challenges associated with large-scale genomic data and the moved on to the traditional and graph based approaches to study microarray data. For the sake of simplicity the graph based methods were not elaborately discussed. Since, microarray and RNA-seq data is the most commonly analyzed genomic data, thus it was covered in detail. With gaining popularity of single-cell based methods, we have introduced the readers to the common methods for clustering these data. In most cases, it is a combination of existing methods like k-means, heirarchical clustering, model based methods, etc.

The survey in the interest of time has curated the commonly used algorithms in the field and does not give a an exhaustive list of all the algorithms. The run-time and clustering efficiency based on evaluation criteria like Jaccard similarity for small, medium and large size cluster can be a good starting point to bench mark all the algorithms. There are not many surveys in the recent decade which have discussed the modifications of these algorithms. The sequencing platforms are being integrative and multimodel, which means along with sequencing data , we have image, antibody, methylation and other data types that need integrative analysis. Hence, building algorithms which can take up these different features, appropriately integrate, remove batch effects and cluster the data can be beneficial for the field.

REFERENCES

[1] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jg Sander. 1999. OPTICS: Ordering Points To Identify the Clustering Structure. (1999).

[2] Chris H Q Ding. 2002. Analysis of gene expression profiles: class discovery and leaf ordering.

[3] Angelo Duò, Mark D. Robinson, and Charlotte Soneson. 2018. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research* 7 (11 2018), 1141. https://doi.org/10.12688/F1000RESEARCH.15666.3

[4] Sofie Van Gassen, Britt Callebaut, Mary J. Van Helden, Bart N. Lambrecht, Piet Demeester, Tom Dhaene, and Yvan Saeys. 2015. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry. Part A : the journal of the International Society for Analytical Cytology* 87 (7 2015), 636–645. Issue 7. https://doi.org/10.1002/CYTO.A.22625

[5] J. C. Gower. 1971. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* 27 (12 1971), 857. Issue 4. https://doi.org/10.2307/2528823

[6] T. Hastie, R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W. C. Chan, D. Botstein, and P. Brown. 2000. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome biology* 1 (2000). Issue 2. https://doi.org/10.1186/GB-2000-1-2-RESEARCH0003

[7] Jeremy J. Jay, John D. Eblen, Yun Zhang, Mikael Benson, Andy D. Perkins, Arnold M. Saxton, Brynn H. Voy, Elissa J. Chesler, and Michael A. Langston. 2012. A systematic comparison of genome-scale clustering algorithms. *BMC bioinformatics* 13 Suppl 10 (6 2012), 1–12. Issue 10. https://doi.org/10.1186/1471-2105-13-S10-S7/FIGURES/4

[8] Daxin Jiang, Chun Tang, and Aidong Zhang. 2004. Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering* 16 (11 2004), 1370–1386. Issue 11. https://doi.org/10.1109/TKDE.2004.68

[9] Vladimir Yu Kiselev, Kristina Kirschner, Michael T. Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N. Natarajan, Wolf Reik, Mauricio Barahona, Anthony R. Green, and Martin Hemberg. 2017. SC3: consensus clustering of single-cell RNA-seq data. *Nature methods* 14 (4 2017), 483–486. Issue 5. https://doi.org/10.1038/NMETH.4236

[10] Jae K. Lee, Paul D. Williams, and Sooyoung Cheon. 2008. Data Mining in Genomics. *Clinics in laboratory medicine* 28 (3 2008), 145. Issue 1. https://doi.org/10.1016/J.CLL.2007.10.010

[11] Peijie Lin, Michael Troup, and Joshua W.K. Ho. 2017. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biology* 18 (3 2017). Issue 1. https://doi.org/10.1186/S13059-017-1188-0

[12] John Quackenbush. 2001. Computational analysis of microarray data. *Nature Reviews Genetics 2001 2:6* 2 (2001), 418–427. Issue 6. https://doi.org/10.1038/35076576

[13] Rahul Satija, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. 2015. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology* 33 (5 2015), 495–502. Issue 5. https://doi.org/10.1038/NBT.3192

[14] Anne Senabouth, Samuel W. Lukowski, Jose Alquicira Hernandez, Stacey B. Andersen, Xin Mei, Quan H. Nguyen, and Joseph E. Powell. 2019. ascend: R package for analysis of single-cell RNA-seq data. *GigaScience* 8 (8 2019). Issue 8. https://doi.org/10.1093/GIGASCIENCE/GIZ087

[15] Mat Soukup, Hyung Jun Cho, and Jae K. Lee. 2005. Robust classification modeling on microarray data using misclassification penalized posterior. *Bioinformatics (Oxford, England)* 21 Suppl 1 (6 2005). Issue SUPPL. 1. https://doi.org/10.1093/BIOINFORMATICS/BTI1020

[16] John D. Storey and Robert Tibshirani. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* 100 (8 2003), 9440. Issue 16. https://doi.org/10.1073/PNAS.1530509100

[17] Ayalew Tefferi, Mark E. Bolander, Stephen M. Ansell, Eric D. Wieben, and Thomas C. Spelsberg. 2002. Primer on Medical Genomics Part III: Microarray Experiments and Data Analysis. *Mayo Clinic Proceedings* 77 (9 2002), 927–940. Issue 9. https://doi.org/10.4065/77.9.927

[18] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* 98 (4 2001), 5116–5121. Issue 9. https://doi.org/10.1073/PNAS.091062498

[19] Dongkuan Xu and Yingjie Tian. 2015. A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science 2015 2:2* 2 (8 2015), 165–193. Issue 2. https://doi.org/10.1007/S40745-015-0040-1

[20] Rui Xu and Donald Wunsch. 2005. Survey of clustering algorithms. *IEEE transactions on neural networks* 16 (5 2005), 645–678. Issue 3. https://doi.org/10.1109/TNN.2005.845141

[21] Aimin Yang, Wei Zhang, Jiahao Wang, Ke Yang, Yang Han, and Limin Zhang. 2020. Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA. *Frontiers in Bioengineering and Biotechnology* 8 (9 2020), 1032. https://doi.org/10.3389/FBIOE.2020.01032/BIBTEX