UNIVERSITATEA ”ALEXANDRU-IOAN CUZA” DIN IAȘI

FACULTATEA DE INFORMATICĂ

LUCRARE DE LICENȚĂ

Aspecte computaționale în biologie

propusă de

Andi Munteanu

Sesiunea: februarie, 2019

Coordonator științific

Conf. Dr. Liviu Ciortuz

UNIVERSITATEA ”ALEXANDRU-IOAN CUZA” DIN IAȘI

FACULTATEA DE INFORMATICĂ

Aspecte computaționale în biologie

Andi Munteanu

Sesiunea: februarie, 2019

Coordonator științific

Conf. Dr. Liviu Ciortuz

Avizat,

Îndrumător lucrare de licență,

Conf. Dr. Liviu Ciortuz.

Data: ..... Semnătura: .....

Declarație privind originalitatea conținutului lucrării de licență

Subsemnatul **Munteanu Andi** domiciliat în **România, jud. Iași, mun. Iași, calea Buzăului, nr. 25, bl. A, et. 5, ap. 45**, născut la data de **01 ianuarie 2018**, identificat prin CNP **1234567891234**, absolvent al Facultății de informatică, **Facultatea de informatică** specializarea **informatică**, promoția 2018, declar pe propria răspundere cunoscând consecințele falsului în declarații în sensul art. 326 din Noul Cod Penal și dispozițiile Legii Educației Naționale nr. 1/2011 art. 143 al. 4 și 5 referitoare la plagiat, că lucrarea de licență cu titlul **Aspecte computaționale în biologie** elaborată sub îndrumarea domnului **Conf. Dr. Liviu Ciortuz**, pe care urmează să o susțin în fața comisiei este originală, îmi aparține și îmi asum conținutul său în întregime.

De asemenea, declar că sunt de acord ca lucrarea mea de licență să fie verificată prin orice modalitate legală pentru confirmarea originalității, consimțind inclusiv la introducerea conținutului ei într-o bază de date în acest scop.

Am luat la cunoștință despre faptul că este interzisă comercializarea de lucrări științifice în vederea facilitării falsificării de către cumpărător a calității de autor al unei lucrări de licență, de diplomă sau de disertație și în acest sens, declar pe proprie răspundere că lucrarea de față nu a fost copiată ci reprezintă rodul cercetării pe care am întreprins-o.

Data: ..... Semnătura: .....

Declarație de consimțământ

Prin prezenta declar că sunt de acord ca lucrarea de licență cu titlul **Aspecte computaționale în biologie**, codul sursă al programelor și celelalte conținuturi (grafice, multimedia, date de test, etc.) care însoțesc această lucrare să fie utilizate în cadrul Facultății de informatică.

De asemenea, sunt de acord ca Facultatea de informatică de la Universitatea "Alexandru-Ioan Cuza" din Iași, să utilizeze, modifice, reproducă și să distribuie în scopuri necomerciale programele-calculator, format executabil și sursă, realizate de mine în cadrul prezentei lucrări de licență.

Absolvent **Andi Munteanu**

Data: ..... Semnătura: .....

Cuprins

Motivație

Diam sit amet nisl suscipit adipiscing bibendum. Aliquet lectus proin nibh nisl condimentum id. Urna duis convallis convallis tellus id interdum velit laoreet. Amet tellus cras adipiscing enim eu turpis egestas pretium aenean. Tortor condimentum lacinia quis vel eros donec ac odio tempor. Volutpat ac tincidunt vitae semper. Urna cursus eget nunc scelerisque viverra mauris in aliquam. Aliquam id diam maecenas ultricies. Molestie a iaculis at erat. Tincidunt nunc pulvinar sapien et ligula ullamcorper malesuada proin. Consequat interdum varius sit amet. Eget est lorem ipsum dolor sit amet consectetur adipiscing. Pharetra diam sit amet nisl suscipit adipiscing bibendum. Maecenas sed enim ut sem viverra aliquet eget sit. Enim blandit volutpat maecenas volutpat blandit aliquam etiam erat velit.

Introducere

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Nunc mattis enim ut tellus elementum sagittis vitae et. Placerat in egestas erat imperdiet sed euismod. Urna id volutpat lacus laoreet non curabitur gravida. Blandit turpis cursus in hac habitasse platea. Eget nunc lobortis mattis aliquam faucibus. Est pellentesque elit ullamcorper dignissim cras tincidunt lobortis feugiat. Viverra maecenas accumsan lacus vel facilisis volutpat est. Non odio euismod lacinia at quis risus sed vulputate odio. Consequat ac felis donec et odio pellentesque diam volutpat commodo. Etiam sit amet nisl purus in. Tortor condimentum lacinia quis vel eros donec. Phasellus egestas tellus rutrum tellus pellentesque eu tincidunt. Aliquam id diam maecenas ultricies mi eget mauris pharetra. Enim eu turpis egestas pretium.

Capitolul 1

Description of methods

This chapter contains informations about the methods used for graph clustering, the sequencing and processing the biological data and eventualy mentions of other works / papers that were focusing on assessing the robustness on changing the seed.

1.1 Graph Clustering

1.1.1 Short intro about what graph clustering is

1.1.2 Why graph clustering instead other traditional methods such as k-means, density based techniques etc

1.1.3 Types of graph clustering

1.1.4 Community detection - Optimizing the quality function

1.1.5 Louvain

The Louvain algorithm [1] is the state-of-the art community detection algorithm that uses an iterative greedy approach. The method can be used to optimize a partition provided by the user, but the default behaviour is to initially assign each node to its own cluster. Each iteration is described by two repeating steps. The first step is to change the partition structure in a greedy manner. For each node the algorithm evaluates whether the change of its label could improve the overall quality or not. If so, the node moves to the cluster that provides the greatest increase of quality. This operations is repeated until no change is longer possible. The second step is shrinking the

graph, meaning each community with a super-node. Thus, the number of nodes of the resulting graph will be the same as the number of clusters that were identified in the first step. The weight of the edges are also recalculated using the sum of inter-cluster weights of the original graph. These two steps are repeated until the partition doesn't change after the first phase.

The algorithm can use multiple runs, when the current iterations takes as starting point the partition that is obtained in the previous one. The algorithm is said to reach convergence if no change is noticed at two consecutive iterations.

Altough it is a greedy algorithm, Louvain proved it can obtain qualitative clusters. Another advantage of the method is computational efficiency: the average time complexity is $O(n \log n)$, where n is the number of nodes.

1.1.6 Louvain refined

1.1.7 SLM

1.1.8 Leiden

1.2 PhenoGraph pipeline

PhenoGraph [2] is a pipeline proposed by Levine et al. to process biological data and obtain a clustering that is interpreted as different cell types. The pipeline is consists of the following steps:

1. dimensionality reduction
2. graph construction
3. graph clustering

Each step will be described in detail in the following sections.

1.2.1 Dimensionality reduction

Given that the human genome contains approximately 25-30000 genes, it is expected that the input data (that is, cells extracted from a tissue from multiple donors) will be highly dimensional. Clustering techinques are highly reliant on calculating distances between points, thus an increased number of dimension will lead to expensive

computations. The solution for this is to reduce the input space such that no information is lost.

One of the most used approach is Principal Component Analysis (PCA) [3], which is a method that uses linear combinations (called principal components) of the initial features to reduce the number of dimensions. This algorithm relies on computing the singular values decomposition, which is a heavy computational task, but several methods of truncating the calculation were developed in order to increase the algorithm's efficiency [4]. To prevent the loss of the original information, the common practice is to use somewhere between 30 and 50 principal components.

Dimensionality reduction can also be performed in a non-linear fashion. Here we mention UMAP [5], an graph-based method that tries to optimize a cross-entropy function in order to create a reduced space that preserves the topology of the original data: the similar points are kept in close proximity, while maintaining the separation between distinct well-defined groups. Compared to the linear methods, UMAP manages to preserve the structure of the original data in only two or three dimensions. This characteristic makes UMAP a more suitable choice when it comes to visualising the data. The downside of the non-linear method that, given its stochastic nature, it can be affected by the value of the random seeds. Usually the effect is presented as slight changes of the topology of the groups or rotations of the representation.

1.2.2 Describing the pipeline

Present the steps that describe the pipeline. (Dimensionality reduction, graph building and graph clustering)

About dimensionality reduction

1.2.3 How to convert matrix data into a graph using kNN

1.2.4 SNN - providing weights using Jaccard Similarity Index

1.3 Element-Centric Similarity

1.3.1 Description about how it works

Describe the intuition behind ECS: the idea of the bipartite graph between points and clusters.

More details about how to calculate ECS. Talk about the affinity matrix and the L1 distance.

1.3.2 Properties, comparison with other clustering metrics

Present some limitation of other clustering metrics such as bias toward cluster sizes, shapes and so on. Perhaps present some comparison figures from the main article.

Present some properties of ECS:

1. the fact that it can be used not only for flat disjoint clusterings, but also for overlapping or hierarchical partitions
2. it overcomes the biases present in the other clustering metrics
3. ECS illustrates the overall similarity between two partitions but also can help in identifying the points where the clustering are not similar

1.3.3 ECC

Talk about how ECC is calculated

1.4 Intro info about biological data and sequencing techniques

Tell about sequencing techniques, how the initial data looks, about cells, genes, what they mean, what is the role and the purpose of the clusters in the biological inter-

pretation.

Capitolul 2

The importance of parameter values in the clustering output

2.1 Monocle and Seurat

Introduce some details about these packages, the language

Capitolul 3

ClustAssess

Amet venenatis urna cursus eget. Quam vulputate dignissim suspendisse in est ante. Proin nibh nisl condimentum id. Egestas maecenas pharetra convallis posuere morbi. Risus viverra adipiscing at in. Vulputate eu scelerisque felis imperdiet. Cras adipiscing enim eu turpis egestas pretium aenean pharetra. In aliquam sem fringilla ut morbi tincidunt augue. Montes nascetur ridiculus mus mauris. Viverra accumsan in nisl nisi scelerisque eu ultrices vitae. In nibh mauris cursus mattis molestie a iaculis. Interdum consectetur libero id faucibus nisl tincidunt eget. Gravida in fermentum et sollicitudin ac orci. Suscipit adipiscing bibendum est ultricies. Etiam non quam lacus suspendisse. Leo urna molestie at elementum eu facilisis sed odio morbi. Egestas congue quisque egestas diam in arcu cursus. Amet consectetur adipiscing elit ut aliquam purus.

3.1 Titlul secțiunii 1

Eros donec ac odio tempor. Facilisi morbi tempus iaculis urna id volutpat. Faucibus in ornare quam viverra orci sagittis eu. Amet tellus cras adipiscing enim eu turpis egestas. Integer feugiat scelerisque varius morbi. Platea dictumst vestibulum rhoncus est pellentesque elit ullamcorper dignissim. Bibendum arcu vitae elementum curabitur. Eu nisl nunc mi ipsum faucibus. Id aliquet lectus proin nibh nisl condimentum id venenatis a. Cras adipiscing enim eu turpis egestas pretium. Quisque non tellus orci ac auctor augue mauris augue. Malesuada pellentesque elit eget gravida cum. Ut lectus arcu bibendum at. Massa id neque aliquam vestibulum morbi blandit. Posuere ac ut consequat semper viverra nam. Viverra adipiscing at in tellus integer feugiat

scelerisque varius morbi. Morbi enim nunc faucibus a pellentesque sit amet porttitor eget. Eu feugiat pretium nibh ipsum consequat nisl vel. Nisl purus in mollis nunc sed.

3.2 Titlul secțiunii 2

Elementum sagittis vitae et leo duis ut diam quam nulla. Purus sit amet volutpat consequat mauris nunc. Tincidunt augue interdum velit euismod in pellentesque massa. Nunc sed augue lacus viverra vitae congue. Porttitor leo a diam sollicitudin. Faucibus pulvinar elementum integer enim. Adipiscing bibendum est ultricies integer quis auctor elit. Blandit aliquam etiam erat velit scelerisque in. A iaculis at erat pellentesque adipiscing commodo elit at. Erat nam at lectus urna duis. Consequat ac felis donec et. Fermentum posuere urna nec tincidunt praesent semper feugiat nibh sed. Proin gravida hendrerit lectus a. Pretium viverra suspendisse potenti nullam ac tortor vitae purus. Arcu cursus euismod quis viverra nibh cras pulvinar mattis. Gravida arcu ac tortor dignissim convallis aenean. Quam nulla porttitor massa id neque aliquam vestibulum morbi. Sed viverra ipsum nunc aliquet. Quis enim lobortis scelerisque fermentum dui faucibus in.

Capitolul 4

Experiments and results

Conclusions, Future Work

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Nunc mattis enim ut tellus elementum sagittis vitae et. Placerat in egestas erat imperdiet sed euismod. Urna id volutpat lacus laoreet non curabitur gravida. Blandit turpis cursus in hac habitasse platea. Eget nunc lobortis mattis aliquam faucibus. Est pellentesque elit ullamcorper dignissim cras tincidunt lobortis feugiat. Viverra maecenas accumsan lacus vel facilisis volutpat est. Non odio euismod lacinia at quis risus sed vulputate odio. Consequat ac felis donec et odio pellentesque diam volutpat commodo. Etiam sit amet nisl purus in. Tortor condimentum lacinia quis vel eros donec. Phasellus egestas tellus rutrum tellus pellentesque eu tincidunt. Aliquam id diam maecenas ultricies mi eget mauris pharetra. Enim eu turpis egestas pretium.

Title: ClustAssess: Tools for Assessing Clustering

Student name: Munteanu Andi

Coordinator: Conf dr. Liviu Ciortuz

The thesis is based on the ClustAssess paper [6], where I contributed as an author and developer of the R package.

Clustering is an unsupervised method that is used to classify and label points based on different similarity metrics. Comparing to the supervised classifiers, clustering algorithms do not require training and using a model and are most suitable when the label of the points are not priorly known and are inferred based on the features that describe them.

Depending on the approach, clustering methods can be divided in multiple categories, such as centroid-based (k-means), density-based (DBSCAN), hierarchical, distribution-based (EM). One approach that gained popularity in the last decades are the graph-based clustering. Some of its advantages is providing flexibility when it comes to the cluster shapes and sizes, or the scalability.

Our thesis focus is set on the community detection techniques, that is graph-based clustering approach that rely on optimizing an objective function (also known in literature as quality function or quality metric, as it tries to evaluate the quality of the clustering by encouraging high density of edges inside clusters and as few intrer-cluster links as possible). This method is intensively used, as it manages to obtain close to optimal results in an efficient time.

Given that many datasets are provided as points displayed on a high-dimensional space, a methodology that enables the use of community detection method on this data is required. Our reference is the PhenoGraph algorithm presented by Levine et. al [2], that establishes a pipeline that firstly applies a dimensionality reduction technique (a linear one such as PCA or non-linear one such as UMAP), then generates a graph using the Nearest Neighbour algorithm. The resulting graph is then used as input for the community detection algorithm.

The PhenoGraph pipeline has been used frequently in several domains. One of them is the downstream analysis of biological data, where the goal is to cluster the cells in multiple groups that will be used to infer some biological conclusions. The input data is usually provided as a matrix that has cells sampled from different donors at

different timepoints on the rows. The columns represent the genes that can be expressed in the cells. Given that the human genome contains approximately 30.000 genes, the input space is a high dimensional one, which makes the PhenoGraph pipeline a suitable choice for processing the data. Also, using graphs to represent data is a more appropiate way to describe the cell-cell interaction.

The current state-of the art of processing the biological data is thus using the PhenoGraph pipeline: for dimensionality reduction, approximate PCA using the Lanczos bidiagonalization method [4] or UMAP [5]; for graph building, the Nearest Neighbour algorithm [7]; for community detection, Louvain [1], Louvain with multilevel refinement [8], Smart Local Moving Algorithm [9] and Leiden [10].

Currently, some of the most popular tools used for analyzing single-cell data are Seurat [11], Monocle [12] and SCANPY [13]. Our thesis makes a comparison between Seurat and Monocle regarding the implementation of the PhenoGraph pipeline. This was motivated by the significant differences between the results of the two packages and the subsequent divergent biological interpretation of the obtained partitions. The question we wanted to answer is whether this differences are caused by computational or biological factors (such as sequencing depth or how the data was pre-processed).

In our thesis we showcase how the divergence was caused by using parameters values that do not match. The conclusion we draw is that tuning the algorithm's parameters is essential in obtaining reproducable results. Given the stochastic nature of the algorithms that are involved in the pipeline, we also noticed how changing the random seed value is a direct factor that affects the clustering output.

The instability caused by random seed was previously identified in several papers that pursued the algorithm modification in order to achieve stability. Such example is kmeans++ [14], where the authors replaced the random initialization of the centroids with assigning probabilities of selection based on the distance to the existing center points. Another example is provided in the clust-perturb algorithm proposed by Stacey et al. [15], where the robustness is evaluated by introducing random noise in the graph.

Our work is focused on providing a pipeline that follows the algorithms involved in the PhenoGraph. The package that we developed its purposed to provide informative plots that would give the user insight about how different parameters impact the number of clusters and the partitioning. Another purpose that we try to achieve is to evaluate and provide insight about parameters configuration that are robust to the

change of seeds. The robustness is determined by using Element Centric Similarity (ECS) [16], a measurement that determines how similar are two clustering of the same data.

Bibliografie

[1] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, October 2008.

[2] Jacob H. Levine, Erin F. Simonds, Sean C. Bendall, Kara L. Davis, El-ad D. Amir, Michelle D. Tadmor, Oren Litvin, Harris G. Fienberg, Astraea Jager, Eli R. Zunder, Rachel Finck, Amanda L. Gedman, Ina Radtke, James R. Downing, Dana Pe'er, and Garry P. Nolan. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, 162(1):184–197, July 2015.

[3] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, 1987. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

[4] James Baglama. Irlba: Fast partial singular value decomposition method. In *Handbook of Big Data*, 2016.

[5] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018. cite arxiv:1802.03426Comment: Reference implementation available at http://github.com/lmcinnes/umap.

[6] Arash Shahsavari, Andi Munteanu, and Irina Mohorianu. Clustassess: tools for assessing the robustness of single-cell clustering. *bioRxiv*, 2022.

[7] Chen Xu and Zhengchang Su. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, 31(12):1974–1980, 02 2015.

[8] Randolf Rotta and Andreas Noack. Multilevel local search algorithms for modularity clustering. *ACM Journal of Experimental Algorithmics*, 16:2.3:2.1–2.3:2.27, July 2011.

[9] Ludo Waltman and Nees Jan van Eck. A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B*, 86(11):471, November 2013.

[10] V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, March 2019.

[11] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, June 2021.

[12] Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J. Steemers, Cole Trapnell, and Jay Shendure. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, February 2019.

[13] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, Feb 2018.

[14] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics.

[15] R. Greg Stacey, Michael A. Skinnider, and Leonard J. Foster. On the robustness of graph-based clustering to random network alterations. *Molecular and Cellular Proteomics*, 20:100002, 2021.

[16] Alexander J. Gates, Ian B. Wood, William P. Hetrick, and Yong-Yeol Ahn. Element-centric clustering comparison unifies overlaps and hierarchy. *Scientific Reports*, 9(1):8574, December 2019.