Unsupervised Multi-Omic Data Fusion: the Neural Graph Learning Network

Unsupervised Multi-Omic Data Fusion: the Neural Graph Learning Network

Pietro Barbiero[1[0000−0003−3155−2564]], Marta Lovino[2[0000−0001−7124−8319]], Mattia Siviero[2[0000−0001−8444−8379]], Gabriele Ciravegna[3[0000−0002−6799−1043]], Vincenzo Randazzo[4[0000−0003−3640−8561]], Elisa Ficarra[2[0000−0002−8061−2124]], and Giansalvo Cirrincione[5,6[0000−0002−2894−4164]]

[1] Cambridge University, Department of Computer Science and Technology,Cambridge, UK
barbiero@tutanota.com
[2] Politecnico di Torino, DAUIN, Turin, Italy
[3] University of Florence, DINFO, Florence, Italy
[4] Politecnico di Torino, DET, Turin, Italy
[5] University of Picardie Jules Verne, Amiens, France
[6] University of South Pacific, Suva, Fiji

Abstract. In recent years, due to the high availability of omic data, data driven biology has greatly expanded. However, the analysis of different data sources is still an open challenge. A few multi-omic approaches have been proposed in literature. However, none of them take into consideration the intrinsic topology of each omic. In this work, an unsupervised learning method based on a deep neural network is proposed. For each omic, a separate network is trained, whose outputs are fused into a single graph; for this purpose, an innovative loss function has been designed to better represent the data cluster manifolds. A graph adjacency matrix is exploited to determine similarities among samples. With this approach, omics having a different number of features are merged into a unique representation. Quantitative and qualitative analyses show that the proposed method has results comparable to the state of the art. The method has a great intrinsic flexibility as it can be customized according to the complexity of the tasks and it has a lot of room for future improvements compared to more fine-tuned methods, opening the way for future research.

Keywords: mRNA, miRNA, lung cancer, multi-omics, SNF, data fusion, neural networks, MLP, unsupervised learning, competitive learning, Kamada-Kawai graph visualization

1 Introduction

In recent years, the development of high throughput techniques for biological data acquisition, like next generation sequencing for DNA and RNA, has significantly increased the availability of raw data, while decreasing the cost by orders of magnitude. For instance, the cost of sequencing a full human genome has fallen from 100 billion

dollars to 1000 dollars in the last 20 years [23]. The availability of this kind of ”omic data” (such as genomic, epigenomic and proteomic data) has remarkably speeded up progress across of biology and medicine. This is also due to emerging cooperative efforts across institutions to build common standardized datasets.

The availability and standardization of data is opening avenues to data driven research, from statistical analysis to supervised and unsupervised machine learning. Supervised learning is limited to the fields where it is possible to obtain accurate labels. One example is the prediction of hard outcomes, like in survival studies [8]. Conversely, unsupervised learning and especially clustering analysis, can lead to the discovery of new classes that may have biological relevance. For instance, clustering of RNA expression data can lead to the discovery of cancer subtypes. [12]. Applying machine learning to single-omic data has produced significant results. mRNA expression data has been successfully used for instance to perform clustering on cancer subtypes or classification based on known sub-types [13]. However, it is limited by the incomplete information carried by single omics. Thus, using multi-omic data integration is of fundamental importance in order to get more accurate analyses and predictions. However, the integration is not trivial and represents an open computational problem.

A solution can be attempted by merging all the features from different omics in a single feature space or performing a consensus clustering among the different input datasets. The former leads, however, to further increase the dimensionality, while the latter is limited in accuracy by the fact that the fusion process is not learnt from the topology of the input spaces. Indeed, multi-omic data integration does not consistently perform better than single omic analysis on the best performing omic [24].

The development of new data fusion techniques is an open research problem. Here the proposed method to address it is a deep learning approach called Neural Graph Learning Fusion (NGL-F).

The paper is organized as follows: Section 2 introduces the background, in particular concerning the problem of applying machine learning to the study of multi-omic data; Section 3 introduces and describes the NGL-F algorithm; Section 4 details the dataset and how the experiments have been performed, comparing the results with those obtained through the Similarity Network Fusion(SNF) algorithm [26], a well-established method for multi-omic data fusion; Section 5, at last, describes the conclusions and the future works.

One of the main contributions of this work is to propose an original neural approach for modeling multi-omic datasets. Compared to the state-of-the-art algorithms, this approach exploits the manifold topology of the input space. The main advantage of this approach is the possibility to extend the algorithm to the case of omics having a different number of samples; this is not possible using the existing techniques, which are not tailored to the problem at hand.

2 Background

Given the greater availability of omic data, thanks to high throughput techniques, data driven biology has greatly expanded with the help of the creation of open databases and the development and improvement of algorithms.

Cooperative effort has led to large scale projects aiming to provide a unified basis for omic data collection and study. Examples are the Ensembl Genome project and the Human Proteome Project, providing respectively a growing data set for the main eukaryotic genes and an attempt to create a map of the protein based molecular architecture of the cell. [15,20]. Similarly, in the medical field, several public databases combine multiple information like omic data, clinical data and histological images, providing the foundation for data driven medical research. Among such projects, the National Cancer Institute Genomic Data Commons (GDC) is a unified data sharing platform for multiple cancer genomic projects. It provides standards for data collection to minimize inconsistencies due to the procedures used. With more than 80'000 samples it constitutes a valuable resource for data driven medical research [18].

Projects like the aforementioned have opened several avenues for computational studies, from statistical analysis to machine learning. The typical problems to be solved are classification and clustering. Clustering problem are of great interest because they allow the discovery of new classes from data beyond human capability. For example, the discovery of new cancer subtypes plays an important role in designing effective therapies that account for resistances. Clustering is an unsupervised learning approach to partitioning sample sets so as to maximize a similarity score among samples in the same subset and minimize it between different subsets [17]. While different computational approaches have produced significant results even with single omics, [13], any omic taken by itself provides an incomplete picture. For example, greater gene expression values for proteincoding genes correlate with higher protein counts for the protein they code for. However, there are regulatory mechanisms that inhibit the translation of mRNA into proteins. One such regulatory element is a small non-coding RNA molecule (miRNA). Thus, combining mRNA and miRNA data should provide a better insight into the cell activity. In general, combining the information from multiple omics is crucial to discover patterns and generate insights at a system level. However, there are significant difficulties to be overcome.

Focusing on multi-omic clustering, different approaches are available. One distinction is between early integration and late integration algorithms: the former unite the features from different omics in a single matrix then perform the clustering; the latter perform clustering separately on the omics then merge the information. Early integration might reveal problematic when the number of samples is much less than the number of features because it increases significantly the dimensionality of the feature space. Late integration is a complex theoretical and computational problem requiring the discovery of new and better algorithms to perform the fusion of the clustering results obtained from each and every omic individually. The difficulties in the use of multi omic data emerge when widely used techniques are benchmarked on real clinical cases an shown not to perform consistently better than single omic data, especially if the comparison is with the best performing omic [24].

One of the state of the art techniques is Similarity Network Fusion (SNF) [27],which starts from the similarity matrices of the original data and creates a consensus through an iterative algorithm: at each step the matrices from individual omics are updated accounting for relevant contributions from the others. This approach has outperformed single-omic studies in some problems such as identification of cancer subtypes and prediction of survival rates when combining mRNA expression, DNA methylation and miRNA expression. The method is simple and fast however it has limitations like requiring to have the same samples across all omics. Although the proposed NGL-F method has been trained on datasets containing the same samples, in principle this is not a strict requirement. Neural networks offer an ample development space: not only they allow to effectively build weighed graphs through strategies such as competitive learning, which then can be merged by accounting for connection strength, but they have built-in tools such as backpropagation to allow each clustering to take into account the information from other omics by introducing a global loss function with coupling terms. An open problem is the determination of well-performing implementations of those coupling terms.

3 The NGL-F neural network

The Neural Graph Learning for data Fusion (NGL-F) is a gradient-based competitive neural network [6], which uncovers topological sample-to-sample relation-ships using multiple data sources. Given two or more types of data for the same set of samples (e.g., patients), NGL-F learns the mutual relationships among samples taking into account such heterogeneous information simultaneously. The output of NGL-F is a set of graphs. For each data set NGL-F aims at finding a graph where nodes represent cluster centroids while edges represent cluster topological properties. Thereafter, the learned topology described by such graphs is used to create the sample adjacency matrix (S). The information contained in the matrix represents all datasets and it can be used to uncover latent patterns among samples. In this sense, the sample adjacency matrix is used to build a unique graph (sample graph) in which nodes represents samples and the edges are derived from S.

NGL-F is composed of a set of dual multi-layer perceptrons (MLPs), one foreach dataset, equipped with a final competitive layer. Weights are estimated by backpropagation. [6]. The activation functions are ReLU for the hidden layers and linear for the output competitive units. The input of each network is a dataset represented as a matrix $X_Z \in \mathbb{R}^{d,n}$, where n is the number of samples and d the number of features. Each MLP provides as output a set of vectors $w \in \mathbb{R}^d$ representing cluster centroids for the input data. For each data source taken into consideration, a multi-layer neural network is instantiated. The architecture of each network can be customized according to the complexity of its own data set (see Fig. 1).

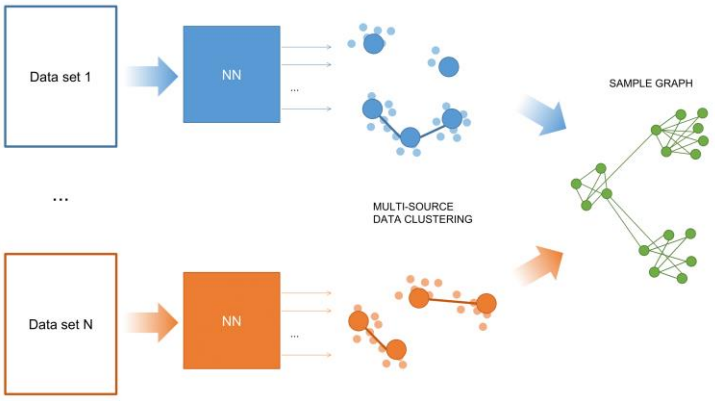

Fig. 1. NGL-F architecture: N datasets are fed in input to NGL-F. For each dataset, a multi-layer perceptron is employed and customized according to dataset complexity. Clustering outputs are at the end combined in order to create a sample graph built from the adjacency matrix S.

The loss function of NGL-F takes into account at the same time the quality of clusters found by each MLP and their underlying topology. The relationships among clusters are modeled using an adjacency matrix E, where E(i, j) represents the number of samples for which w_i and w_j are the two closest centroids. The higher E(i, j), the more their respective clusters are related. The matrix E represents a graph on the neural network, where the nodes are the neurons and the edges are inter-neuron connections. These links represent the topology of the input data. The loss function of each MLP is composed of four terms taking into account inter and intra-cluster distances, quantization error and parsimony in representing the underlying topology:

$$\mathcal{L} = \frac{max_k\, d_{\{intra\}}(C_k)}{max_{\{i,j\}} d_{\{inter\}}(C_i, C_j)} + Q + ||\mathrm{E}|| \tag{1}$$

where $d_{\{intra\}}(C_k)$ is the intra-cluster distance, $d_{\{inter\}}(C_i, C_j)$ the inter-cluster distance, and Q the quantization error. The complete diameter distance is used as an intra-cluster quality index, representing the distance between the two remotest samples belonging to the same cluster:

$$d_{\{intra\}}(C_i) = \max_{x,y \in C_i} d(x, y) \tag{2}$$

The single linkage distance, representing the closest distance between two samples belonging to two different clusters, is used to model inter-cluster distance:

$$d_{\{inter\}}(C_i) = \min_{x \in C_i, y \in C_j} d(x, y) \tag{3}$$

The quantization error is computed as the norm of the distances between cluster centroids (w_i) and cluster points (C_i):

$$Q = \|d(w_i, x)\|_2 \qquad \forall x \in C_i \tag{4}$$

The NGL-F loss function is the linear combination of MLPs' losses:

$$\mathcal{L} = \sum_z \mathcal{L}_Z \tag{5}$$

Once all networks terminate the training procedure, the resulting clusters are analyzed. For each data set, two samples are considered near to each other in case they belong to the same cluster; far from each other in case they belong to different clusters. A sample adjacency matrix *S* is then computed as follow:

$$S(i,j) = \sum_{d=1}^{n} near_d(i,j) \quad (6)$$

where $near_d(i,j)$ is a boolean function calculating the proximity of the samples as previously explained and n is the number of data set taken into consideration. This matrix is the result of the fusion process. Its quality can be analyzed and compared to other methods in different ways, as it will be shown in the next section.

4 Experiments

Data are downloaded from the portal of the NIH Genomic Data Commons [22]and are collected in tabular form, resulting in a mRNA and a miRNA transcriptome profiling matrix.

The mRNA matrix consists in raw counts gene expression values [4]. A higher value represents, for protein coding genes, a greater amount of protein produced. This is true unless regulatory mechanisms inhibit the translation of the mRNA.

The miRNA matrix consists in raw counts miRNA values [9]. As miRNA inhibits the translation of mRNA, a higher expression value corresponds to a lower presence of the proteins related to that sequence.

The data was preprocessed as follows:

- For the mRNA matrix, the genes with an expression value equal to zero across all the samples were deleted. Then the normalization was performed through a variance stabilizing transformation [16] and only protein coding genes were selected. This resulted in 17682 genes for which the expression value is reported.
- For the miRNA matrix, the sequences with zero expression value across the samples were deleted and the matrix was normalized through DESeq2 [21]. The final values were obtained as $\log_2(exprValue + 1)$ [3].

The patients for which either the mRNA or the miRNA data was missing were deleted from the matrices. This resulted in 1248 miRNA and mRNA sequences for which the expression value is reported. This deletion is not a strict requirement in general for NGL-F but it is necessary to compare it with SNF and taken as requirement for this specific implementation.

Data samples come from either healthy or cancerous lung tissue belonging to two types: Lung Adenocarcinoma (LUAD) or Lung Squamous cells Carcinoma (LUSC). The healthy tissue has been taken from non-tumoral tissue samples usually close to the position of the tumor. Data was acquired from three projects: TCGA-LUAD [25] and CPTAC-3, with samples from adenocarcinoma patients, and TCGA-LUSC, with samples from squamous cells carcinoma patients. Overall this resulted in six different annotations all reported as the name of the project followed by either the ”tumoral” or ”healthy” annotation.

All the code for the experiments has been implemented in Python 3, relying upon open-source libraries [1,14]. All the experiments have been run on the same machine: IntelR©CoreTMi7-8750H 6-Core Processor at 2.20 GHz equipped with8 GB RAM.

The two datasets previously described are fed as input to the NGL-F algorithm. The structure of the networks employed in this paper is reported in Fig.2. NGL-F is a single neural network that employs a set of dual multi-layer perceptrons, one for each analyzed omic. The use of dual networks is justified given the high-dimensionality of the data sources [2,6,7]. The number of features may vary between different omic and it is maintained through the layers, as dual networks are trained on the transposed matrix [6]. In this way, output nodes preserve input dimensionality and can be used as cluster centroids for each input matrix. In this implementation, the only requirement is on the number of samples (1248) that needs to be identical among the omics. As mentioned in Sec. 3, the fusion process consists in the creation of a unique sample adjacency matrix that takes into consideration the information extracted from every omic data. In order to compare the results of the proposed method, the experiment was repeated by using the SNF algorithm [26].

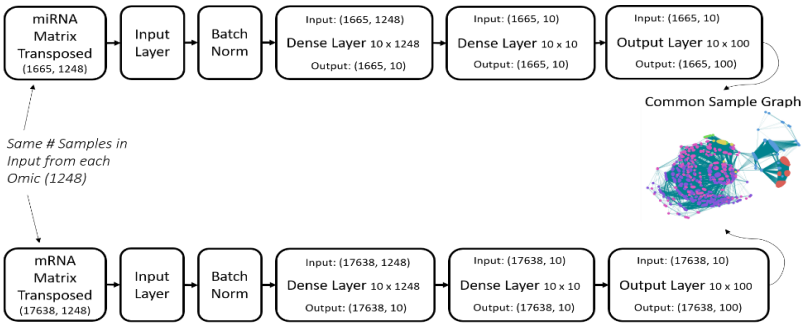

Fig. 2. NGL-F network architecture as used in the experiments. Between brackets the dimensionality of input/output data of each layer are reported. Regarding the matrices, the dimensions are defined as features x samples since the matrix is transposed. Next to each dense and output gclayer, instead, is reported the dimensionality of the associated weight matrix. Also, it should be noticed the different dimensionality of the two input sources, miRNA (top) and mRNA (bottom) maintained through the layers.

The adjacency matrix built by both methods are depicted in Fig. 3. Observing the two plots, the results are pretty similar with both methods capable of identifying similarities among data. This is a first important result as it shows the quality of the fusion process carried out by the proposed method when com-pared to a state-of-the-art algorithm.

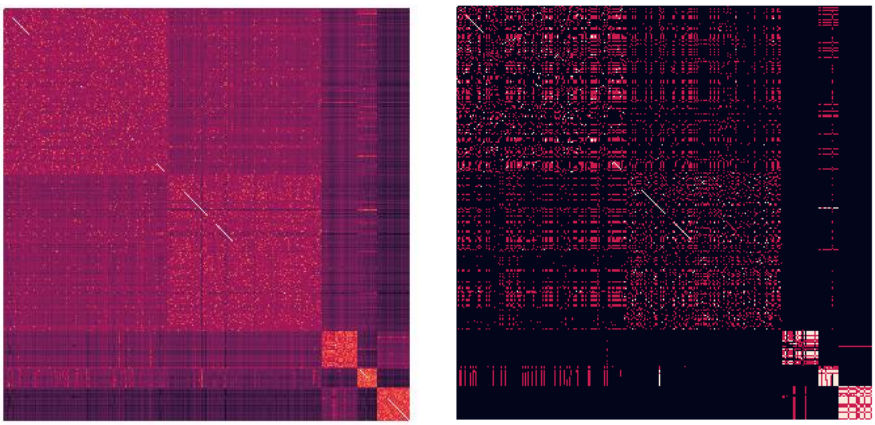

Fig. 3. Adjacency matrix of the sample using (left) SNF and (right) NGL-F algorithms

In order to better analyze this result, it was decided to plot the sample adjacency matrix through the Kamada-Kawai path-length algorithm [19]. This algorithm is a force-directed graph drawing method that can be used to visualize undirected graphs in a two-dimensional space. The main characteristics of this class of algorithms is that edges are displayed in such a way that the number of crossings is the lowest possible. In the two plots of Fig. 4, it is clear that the number of connections found by SNF is redundant: even isolated samples as the LUAD tumoral ones on the top and left edges are connected with many other samples. Conversely, NGL-F better identifies outliers as it can be seen with the tumoral CPTAC3 on the top right corner. However, the sample adjacency matrix plot produced by SNF, better separates LUAD from LUSC tumoral data, while in the plot concerning NGL-F the samples belonging to the two classes are quite confused.

At last, the quality of the proposed algorithm is validated through a comparison of the spectral clustering executed on the two adjacency matrices. In Fig.5, the quality of the clusters of the grouped samples can be appreciated. More precisely, a harmonic mean of purity and efficiency of the clusters is computed according to the class of the samples belonging to each cluster. Both clustering techniques are capable to precisely identify CPTAC3 healthy samples, grouped in the C5 cluster. Also, CPTAC3 tumor samples are mostly collected in a single cluster, C4; however, the adjacency matrix produced by SNF seems to better separate these samples as the corresponding cluster quality is higher. Instead, samples belonging to LUAD and LUSC (both tumor and healthy) seem to be more difficult to identify. Indeed, for both tissues, tumor samples are collected together in the C0 and C2 clusters for SNF and C0 and C1 clusters for NGL-F. At last, the few LUAD and LUSC healthy samples are mostly placed in C3cluster for NGL-F, while they are split among all the clusters in the case of SNF.

Summing up, the results produced by the two algorithms are very similar. It is worth pointing out the importance of this result, as NGL-F is a completely new algorithm and it is based on a recent neural theory [6]. Compared to state-of-the-art methods, the neural network structure of NGL-F shows a higher flexibility and can be easily extended to omics with different number of samples. Future works may include the improvement

of the loss function taking account cluster densities [11] and the development of incremental, hierarchical [10], and biclustering [5] versions of NGL-F.

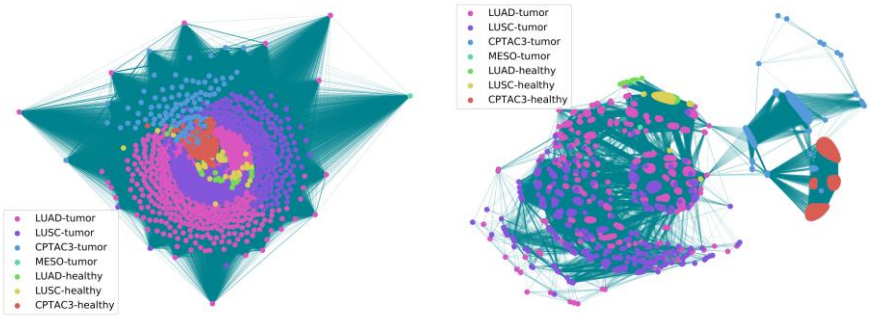

Fig. 4. Graph of the sample adjacency matrix through the Kamada-Kawai path-length algorithm.

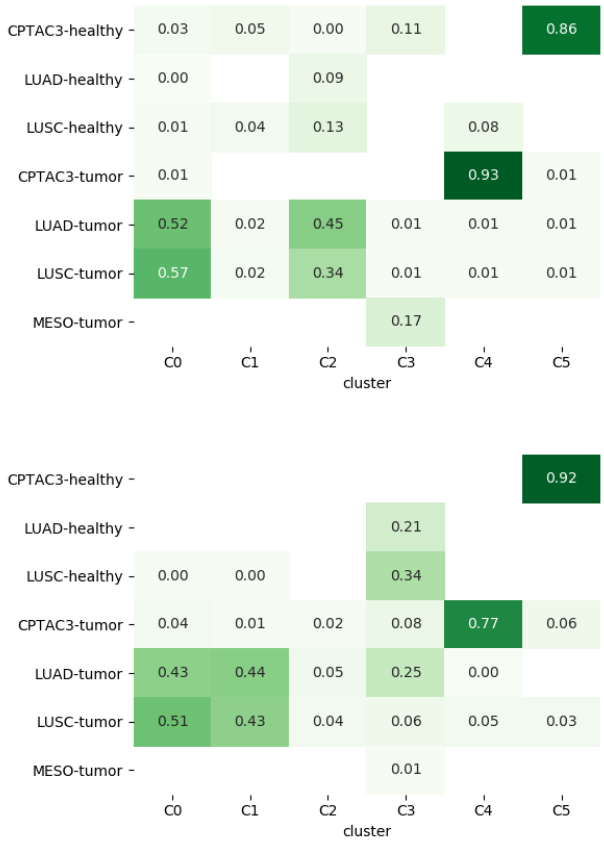

Fig. 5. Harmonic mean of cluster efficiency and purity computed on the spectral clusters, computed on the adjacency matrix produced by SNF (top) and NGL-F (bottom)algorithms

5 Conclusions

Since the interpretation of data coming from multiple data sources is still an open and challenging problem, some multi-omic approaches have been recently proposed. However, these methods do not take into account the intrinsic topology of each omic. Therefore, NGL-F has been designed to tackle this issue. It is an unsupervised deep learning neural network endowed with an original final layer which is competitive, because of the choice of the loss function. Indeed, it takes into account both the quantization and the clustering and the onset of the edges. The training procedure is repeated for all input datasets generating each a network of centroids to which the samples are assigned in a competitive fashion, with criteria for creating and decaying connections between the centroids themselves. The final outcome is a connected graph for each input which is merged to obtain the final graph from which the clusters are derived. Experimental results show its competitiveness with state-of-the-art algorithms; however, they are more flexible in the sense that several kinds of layers can be employed and more than two input sources can be fed simultaneously. Hence, the proposed algorithm is suitable for a wider range of applications.

Future work will deal with the implementation of convolutional layers into the neural architecture and with a deeper analysis of the loss function. A shallow version of the network, which underlines both the competitive aspect of the approach and the topology of the data by the edges, is under study. It will be applied not only to few omics and also to non-biological data.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: 12th {USENIX}Symposium on Operating Systems Design and Implementation ({OSDI}16). pp. 265–283 (2016)
2. Altman, N., Krzywinski, M.: The curse (s) of dimensionality. Nat Methods15(6), 399–400 (2018)
3. Anders, S., Huber, W.: Differential expression of rna-seq data at the gene level – the deseq package. Heidelberg, Germany: European Molecular Biology Laboratory (EMBL)10, f1000research (2012)
4. Anders, S., Pyl, P.T., Huber, W.: Htseq—a python framework to work with high-throughput sequencing data. Bioinformatics31(2), 166–169 (2015)
5. Barbiero, P., Bertotti, A., Ciravegna, G., Cirrincione, G., Cirrincione, M., Piccolo, E.: Neural biclustering in gene expression analysis. In: International Conference on Computational Science and Computational Intelligence (2017)
6. Barbiero, P., Ciravegna, G., Randazzo, V., Cirrincione, G.: Topological gradient-based competitive learning (2020)
7. Barbiero, P., Squillero, G., Tonda, A.: Modeling generalization in machine learning: A methodological and computational study (2020)
8. Chaudhary, K., Poirion, O.B., Lu, L., Garmire, L.X.: Deep learning–based multi-omics integration robustly predicts survival in liver cancer. Clinical Cancer Research 24(6), 1248–1259 (2018)

9. Chu, A., Robertson, G., Brooks, D., Mungall, A.J., Birol, I., Coope, R., Ma, Y., Jones, S., Marra, M.A.: Large-scale profiling of micrornas for the cancer genome atlas. Nucleic acids research44(1), e3–e3 (2016)
10. Cirrincione, G., Ciravegna, G., Barbiero, P., Randazzo, V., Pasero, E.: The gh-exin neural network for hierarchical clustering. Neural Networks121, 57–73 (2020)
11. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. vol. 96, pp.226–231 (1996)
12. Gao, S., Qiu, Z., Song, Y., Mo, C., Tan, W., Chen, Q., Liu, D., Chen, M., Zhou, H.: Unsupervised clustering reveals new prostate cancer subtypes. Translational Cancer Research6(3), 561–572 (2017)
13. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P.,Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439), 531–537 (1999)
14. Hagberg, A., Swart, P., S Chult, D.: Exploring network structure, dynamics, and function using networkx. Tech. rep., Los Alamos National Lab.(LANL), LosAlamos, NM (United States) (2008)
15. Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T.,Cuff, J., Curwen, V., Down, T., et al.: The ensembl genome database project. Nucleic acids research30(1), 38–41 (2002)
16. Huber, W., Von Heydebreck, A., S¨ultmann, H., Poustka, A., Vingron, M.: Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics18(suppl1), S96–S104 (2002)
17. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM computing surveys (CSUR)31(3), 264–323 (1999)
18. Jensen, M.A., Ferretti, V., Grossman, R.L., Staudt, L.M.: The nci genomic data commons as an engine for precision medicine. Blood, The Journal of the American Society of Hematology 130(4), 453–459 (2017)
19. Kamada, T., Kawai, S.: An algorithm for drawing general undirected graphs. Information Processing Letters31(1), 7 – 15 (1989). https://doi.org/https://doi.org/10.1016/0020-0190(89)90102-6,http://www.sciencedirect.com/science/article/pii/0020019089901026
20. Legrain, P., Aebersold, R., Archakov, A., Bairoch, A., Bala, K., Beretta, L., Bergeron, J., Borchers, C.H., Corthals, G.L., Costello, C.E., et al.: The human proteome project: current state and future direction. Molecular & cellular proteomics10(7)(2011)
21. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome biology15(12), 550 (2014)
22. National Cancer Institute: Gdc data portal, https://portal.gdc.cancer.gov/, last accessed on 2020-06-14
23. National Human Genome Research Institute: The cost of sequencing a human genome, https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost, last accessed on 2020-06-14
24. Rappoport, N., Shamir, R.: Multi-omic and multi-view clustering algorithms: re-view and cancer benchmark. Nucleic acids research 46(20), 10546–10562 (2018)
25. Tomczak, K., Czerwinska, P., Wiznerowicz, M.: The cancer genome atlas (tcga): an immeasurable source of knowledge. Contemporary oncology 19(1A), A68 (2015)
26. Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., Goldenberg, A.: Similarity network fusion for aggregating data types on a genomic scale. Nature methods 11(3), 333 (2014)

27. Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., Goldenberg, A.: Similarity network fusion for aggregating data types on a genomic scale. Nature methods11(3), 333 (2014)