Rotation Clustering: A Consensus Clustering Approach to Cluster Gene Expression Data

Paola Galdi^(✉), Angela Serra, and Roberto Tagliaferri

NeuRoNe Lab, DISA-MIS, University of Salerno,
via Giovanni Paolo II 132, 84084 Fisciano, SA, Italy
{pgaldi,aserra,robtag}@unisa.it

Abstract. In this work we present Rotation clustering, a novel method for consensus clustering inspired by the classifier ensemble model Rotation Forest. We demonstrate the effectiveness of our method in a real world application, the identification of enriched gene sets in a TCGA dataset derived from a clinical study on Glioblastoma multiforme.

The proposed approach is compared with a classical clustering algorithm and with two other consensus methods. Our results show that this method has been effective in finding significant gene groups that show a common behaviour in terms of expression patterns.

Keywords: Clustering · Consensus clustering · Rotation Forest · Gene set enrichment · Pathways · Glioblastoma

1 Introduction

Technological advances lead to a huge increase in the number of technologies available to produce *omics* data such as gene expression, RNA expression (RNA), microRNA expression (miRNA), protein expression etc.

Nowadays, especially for microarray gene expression technology, the greatest effort no longer consists in the production of data, but in their interpretation to gain insights into biological mechanisms.

Microarray gene expression data allow to quantify the expression of thousands of genes across hundreds of samples under different conditions [4]. Here the main idea is that genes with similar expression patterns can have a relation in functional pathways or be part of a co-regulation system. This analysis is usually performed with exploratory techniques such as cluster analysis [7].

Clustering is an unsupervised technique used in data analysis to detect natural groups in data without making any assumption about their internal structure. There are two main reasons for choosing such an approach, that are (1) to try to confirm a hypothesis (e.g. about latent classes of objects) (2) to uncover previously unknown relationships among data points.

P. Galdi and A. Serra—Equal contribution.

A. Petrosino et al. (Eds.): WILF 2016, LNAI 10147, pp. 229–238, 2017.
DOI: 10.1007/978-3-319-52962-2_20

Clustering has been successfully applied in bioinformatics in cancer subtyping [5,19,20,23] and in identifying groups of genes that show a similar behaviour [10,14,16].

However, one of the issues of full data-driven approaches (as opposed to hypothesis-driven approaches) is the risk of modelling noise or uninteresting properties w.r.t. the problem domain.

This is even more true for microarray gene expression data, since they are characterized by intrinsic noise, due to the high dynamics of the studied systems. Moreover, data have a background noise related to mechanical tools used to perform the analyses. For this reason, robust clustering techniques, such as consensus clustering, have been applied in gene expression clustering [9,17].

Consensus clustering has been devised as a method to deal with these issues by combining multiple clustering solutions in a new partition [22]. The main idea is to exploit the differences among base solutions to infer new information and discard results that might have been affected by the presence of noise or by intrinsic flaws of the chosen clustering algorithm. Being a consensus solution supported by the agreement of several base clusterings, not only it is more stable and robust to overfitting, but it also guarantees a higher degree of confidence in the results.

The ratio behind consensus clustering is analogous to that of classifier ensemble, where multiple "weak" classifiers are combined to obtain better performances [3]. Previous results have shown how diversity in the initial clustering solutions can lead to an improvement of the quality of the final consensus clustering [15]. For instance, in [8] the authors have investigated the relation between clustering accuracy (w.r.t. known classes of points) and the average normalized mutual information between pairs of partitions.

Since then many concepts have been borrowed from the classifier ensemble literature, such as subsampling or projections of the original data to promote diversity in the base partitions. Following this idea, here we propose a novel consensus clustering technique inspired by the Rotation Forest classifier [18] called Rotation Clustering.

2 Materials and Methods

2.1 Consensus Clustering

Consensus techniques are characterized by how the diversity among base solutions is generated and how the agreement among clusterings is quantified. In the following, two representative examples of consensus clustering methods are presented.

The approach by Monti et al. [17] generates multiple perturbed versions of the original data by computing random subsamples of the input matrix. Then a consensus (or co-association) matrix $M \in \mathcal{R}^{n \times n}$ (where n is the number of data objects) is built, where each entry $M(i, j)$ is the count of how many times items i and j were assigned to the same cluster across different partitions, normalized by the number of times that the two objects were present in the same subsample.

The approach by Bertoni and Valentini [1] uses random projections to build the perturbed versions of the input data. Random projections [2] are based on two main results. The first is the *Johnson-Lindenstrauss lemma* [13], that can be summarized as follows: if points in a vector space are projected onto a randomly selected subspace of suitably high dimension, then the distances between the points are approximately preserved. The second is the *Hecht-Nielsen lemma* [12], that states that "in a high-dimensional space, there exists a much larger number of almost orthogonal than orthogonal directions. Thus, vectors having random directions might be sufficiently close to orthogonal" [2]. Starting from these premises, the idea is to project the original d-dimensional data set to different k-dimensional ($k << d$) subspaces using random matrices whose elements are Gaussian distributed; a clustering is then executed on each subspace. Both rotation clustering and the approach based on random projections use the co-association matrix to measure the level of consensus. The final clustering can be obtained using the consensus matrix as a similarity matrix to be given as input to a hierarchical clustering algorithm.

In all the experiments the base solutions to be combined are generated with a single execution of the k-means algorithm with random initialization of the initial centroids.

2.2 Rotation Forest

Rotation Forest [18] is a classifier ensemble method that trains several base classifiers in the following way: starting from the training data matrix X (n samples $\times$ d features), the feature set is partitioned into K subsets, then from each of the submatrices X_i (for $i = 1, \ldots, K$) extracted from X selecting only one of the K subsets of features, a random subset of classes is eliminated and the remaining items are subsampled to obtain a sample size of, say, 75% of the original number of objects; PCA is applied on each one of the resulting matrices X_i and the computed coefficients are arranged in matrices C_i. All the components extracted by PCA are retained to not disrupt discriminatory information that might lie in the last components. The C_i matrices are then combined to build a sparse rotation matrix R which is arranged in such a way that its columns match the order of the original feature set; finally, the classifier is trained using XR as the training set.

This method has been proven to be able to outperform well established techniques in the classifier ensemble literature [18].

2.3 Rotation Clustering

Rotation Clustering follows the same steps of the Rotation Forest for building the input matrix except for what concerns the removal of a subset of classes since no prior information is available (see Algorithm 1 for the pseudo-code of the algorithm). Each of the input matrices for the base clustering is generated in the following way: first features are split randomly in subsets and for each subset a submatrix is built that contains only the features in the subset and a random subsample of the original data items; PCA is applied on each of the submatrices and a rotation matrix is built by combining the coefficients of all the principal components of each submatrix; finally the original data matrix is rotated using the obtained rotation matrix. Once the base clusterings are computed, a pair-wise co-association matrix is built counting how many times each pair of data objects was assigned to the same cluster across base solutions. The final consensus solution is the result of a hierarchical clustering algorithm applied to the co-association matrix.

Algorithm 1. Rotation clustering

Input:

- data matrix $X \in \mathcal{R}^{n \times d}$ where n is the no. of samples and d is the no. of features.
- the number M of base clusterings to generate
- the number K of feature subsets

```
for $i = 1, \ldots, M$ do
    Split the feature set in $K$ subsets of size $\lfloor \frac{n}{K} \rfloor$
    for $j = 1, \ldots, K$ do
        $X_{i,j} \leftarrow$ select from $X$ the $j - th$ subset of features $j$
        $X'_{i,j} \leftarrow$ select a subsample of items from matrix $X_{i,j}$
        $C_{i,j} \leftarrow$ apply PCA on matrix $X'_{i,j}$                ▷ PCA coefficients
    end for
    Arrange the $C_{i,j}$ in a rotation matrix $R_i$
    $X'_i \leftarrow XR_i$                                         ▷ Rotate input matrix
    $cls_i \leftarrow$ apply clustering algorithm on $X'_i$
end for
$final \leftarrow$ combine the $cls_i$ in a consensus solution
return $final$
```

3 Experimental Setup and Validation

We compared the results obtained with the proposed method with those produced by the best of 100 runs of k-means with random initialization of the centroids and by two consensus techniques found in literature: the former builds a co-association matrix starting from subsamples of the original data; the latter is based on random projections (see Sect. 2.1).

The base solutions to be combined by each of the three consensus algorithms are generated by single runs of the k-means algorithm. The choice of using the same clustering algorithm across experiments is motivated by our dual goal of (1) comparing the performance of consensus methods versus a simple run of a clustering algorithm and (2) assessing the efficiency of our method compared to that of existing techniques. All clustering algorithms used Pearson's correlation coefficient as similarity measure between genes, since the goal is to identify gene sets that show a similar behaviour in terms of expression patterns.

The validation process considers two fundamental aspects of the gene clustering problems. Firstly, cluster analysis is a complex task and the results can change based on the number K of clusters selected as input parameter [11]. Therefore, each clustering algorithm was executed with different values of K (50, 100, 150, 200 and 250) and the performance was evaluated according to the index proposed in [19]: $ClVal = \frac{1}{4}\left(\frac{IC+1}{2} + 1 - \frac{EC+1}{2} + (1-S) + CG\right)$. This index takes into account the average sample correlations inside each cluster (IC), the average sample correlation of the least similar objects for each pair of clusters (EC), the number of singletons (S) and the compression gain (CG). Its range is between 0 and 1, the higher the value, the better is the clustering result. Secondly, once the cluster analysis is performed on the gene expression dataset, the gene group needs to be interpreted from a biological perspective. Therefore, after selecting the best value of K, an enrichment set analysis was performed between all the clustering algorithms. The aim of the gene set enrichment is to evaluate microarray data at the level of gene sets [21]. Gene sets are defined based on *a priori* knowledge and, usually, they are gene sets with similar characteristics and behaviour. This is used to evaluate if the genes in a specific cluster have an homogeneous biological behaviour. This kind of analysis compares the clusters obtained with different methods by counting the number of gene sets enriched by each cluster and evaluating the gene ratio. For each pair of cluster and pathway, the gene ratio measures the proportion of genes in the cluster that are also included in the pathway. The best method is the one that, overall its clusters, has a higher number of enriched sets with the highest gene ratio. The analysis was performed by using the KEGG pathways from the Kyoto Encyclopedia of Genes and Genomes and the *compareCluster* function from the *clusterProfiler* R package. The last part of validation consists in verifying how many gene sets associated with specific diseases are identified by the used clustering algorithms. Known association between pathways and Glioblastoma were downloaded from the Comparative Toxicogenomics Database (CTD - http://ctdbase.org/) [6].

Dataset. Experiments have been performed on a real gene expression dataset, related to a Glioblastoma multiforme study. The dataset was accesses through the TCGA website (https://tcga-data.nci.nih.gov/tcga/ - Glioblastoma multiforme [GBM]) and publicly available gene expression data (level 3) were downloaded from 167 samples. As a further preprocessing step, features with low variance were eliminated and batch effect removal was performed with the *comBat* method in the *sva* R package).

4 Results and Discussion

We developed a new consensus clustering algorithm called Rotation Clustering. We applied this method to the problem of clustering gene expression data. Analyses were performed on a real gene expression dataset from TCGA repository with 2408 genes and 167 samples.

We clustered the genes with our method and we compared the results with other three classical clustering algorithms: the Kmeans clustering, the consensus clustering proposed by Monti et al. [17] and the random projection techniques proposed by Bertoni and Valentini [1].

Clustering algorithms give different results depending on the input parameters. To avoid this problem, the analyses were performed varying the number of clusters to be retrieved (K). The algorithm results were then evaluated according to the $ClVal$ measure proposed by [19]. Figure 1(a) shows that the best value for the parameter K is 50, in fact its score (light blue line) is the highest across the four clustering methods.

Moreover, in order to characterize the biological meaning of the obtained clusters, we performed an enrichment analysis with respect to the KEGG pathways. Figure 1(b) shows that the number of gene sets enriched by the clusters obtained from the Rotation clustering is the highest compared to the other methods when the parameter K assumes the values of 50, 100 and 150. On the other side, when the K value assumes values of 200 and 250, the best algorithm is the one proposed by Monti et al. To obtain a summary assessment of the algorithms we merged these 5 rankings with the Borda count method, implemented in the *TopKLists* R package. The final rating of the algorithm shows that the Rotation method is at the top, followed by the Random Projection approach, then the consensus clustering proposed by Monti et al. and in the last position the k-means algorithm. An important remark is that all the approaches based on consensus clustering give better results compared to the k-means. This justifies the higher computational effort that they require since they give more stable and less noisy results. We further investigated the obtained clusters with respect to the gene ratio. Here we report the results only for $K = 50$. Table 1 shows the quantiles of the distribution of the gene ratio for each clustering algorithm. The Rotation clustering, in addition to being the one with the highest count of gene sets, also reaches the highest value for the gene ratio. That means that there exists at least one cluster of genes in its solution that is completely included in a gene set.

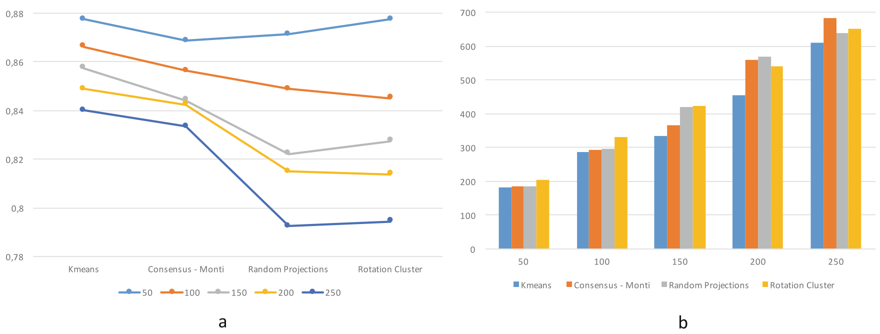

Fig. 1. (a) Evaluation or the $ClVal$ index when changing the parameter K. The figure shows the evaluation index for the parameter K for all the clustering algorithms and the different values of K. (b) Number of enriched KEGG sets. The figure shows the number of enriched KEGG sets for each algorithm and for each k. (Color figure online)

Table 1. Gene ratio

	k-means	Monti-consensus	Random projection	Rotation clustering
0%	0.05	0.04	0.04	0.03
25%	0.11	0.10	0.11	0.10
50%	0.14	0.14	0.17	0.14
75%	0.19	0.20	0.25	0.24
100%	0.42	1.00	0.57	1.00

Gene Ratio with K = 50

More details on the enriched pathway are shown in Fig. 2. The figure shows only the clusters that enrich at least one pathway. The clusters on the x-axis are ordered by gene-ratio. To have more insights into the biological meaning of the problem, we also checked if the pathways enriched from the four different

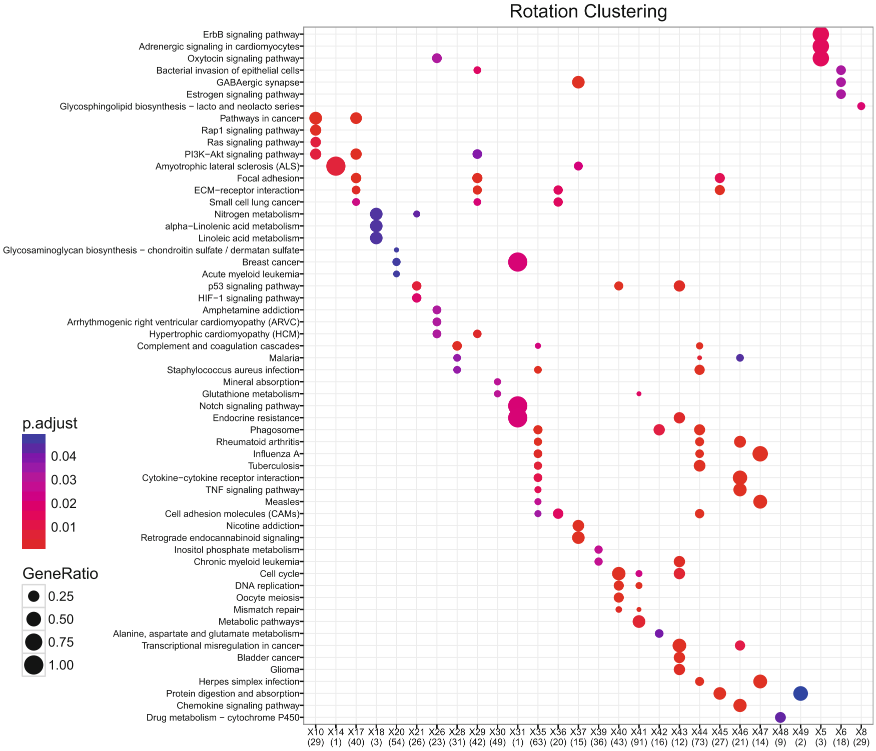

Fig. 2. Enrichment result for Rotation clustering (K = 50). The figure reports the enrichment analysis results for the clusters obtained with the Rotation clustering algorithm with $K = 50$ as input. Clusters with a significant p-value are reported on the x-axis, while pathways are reported on the y-axis. Colour and size of the bubbles indicate the p-value and gene ratio respectively. (Color figure online)

solutions are known to be, somehow, related to Glioblastoma. To do this, we downloaded a list of known pathways related to Glioblastoma from the CTD dataset [6]. This list contains 219 KEGG pathways. We then counted how many of these sets are in our solutions. Figure 3 shows the Venn diagram representing the number of pathways (related to glioblastoma) shared among the four algorithms. As we can see from the diagram, the Rotation Clustering solution is the one with the highest number of enriched pathways. Particularly, k-means is able to retrieve 46 pathways, the Monti's consensus method retrieves 48 pathways, the Random Projection method retrieves 49 of them and the Rotation clustering enriches 51 clusters. Most of the pathways of each of the three consensus methodologies are shared with the ones of the k-means. This can be due to the fact that the consensus algorithms are all based on the k-means, but the number of enriched sets increases when the variability imposed by the consensus methods to the data grows. In fact, while the Monti's method induces a relatively smaller perturbation by subsampling the data but preserving the original distribution of points in space, random projection and rotation also perturb the relative distance among points, probably allowing for previously unobserved relationships to emerge.

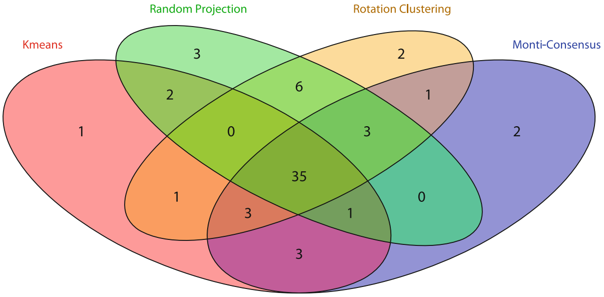

Fig. 3. Number of KEGG gene sets associated to Glioblastoma that give a significant p-value for each clustering technique.

5 Conclusion

In this work we presented a new consensus clustering method called Rotation clustering, that is based on the same idea of the Rotation Forest classifier. We successfully applied this method to a real gene expression clustering problem, related to a clinical study about patients affected by Glioblastoma. We validated our results with respect to both the structure of the clusters and the prior biological knowledge. We also compared the new method with a classical clustering and other consensus-based methodologies. We can conclude that this is an effective method for clustering noisy data because it gives stable and reliable results that resemble the known biological information. The reasons of the efficacy of this approach may be found in how diverse are the base clusterings that are combined in the final consensus solution, and this aspect will be further investigated in future work. However, diversity alone cannot guarantee the quality of the results, especially if we try to merge poor or incompatible clusterings. One possible way to overcome this problem might be merging only partitions that are sufficiently similar, thus adding a meta-clustering step to the consensus framework.

Acknowledgments. We would like to thank Teresa Savino and Luca Puglia for the helpful discussions.

References

1. Bertoni, A., Valentini, G.: Random projections for assessing gene expression cluster stability. In: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, vol. 1, pp. 149–154. IEEE (2005)
2. Bingham, E., Mannila, H.: Random projection in dimensionality reduction: applications to image and text data. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 245–250. ACM (2001)
3. Brown, G.: Ensemble learning. In: Sammut, C., Webb, G.I. (eds.) Encyclopedia of Machine Learning, pp. 312–320. Springer, Heidelberg (2011)
4. Brown, P.O., Botstein, D.: Exploring the new world of the genome with DNA microarrays. Nat. Genet. **21**, 33–37 (1999). http://www.nature.com/doifinder/10.1038/4462
5. Chang, H.Y., Nuyten, D.S., Sneddon, J.B., Hastie, T., Tibshirani, R., Sørlie, T., Dai, H., He, Y.D., van't Veer, L.J., Bartelink, H., et al.: Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc. Nat. Acad. Sci. US Am. **102**(10), 3738–3743 (2005)
6. Davis, A.P., King, B.L., Mockus, S., Murphy, C.G., Saraceni-Richards, C., Rosenstein, M., Wiegers, T., Mattingly, C.J.: The comparative toxicogenomics database: update 2011. Nucleic Acids Res. **39**(suppl 1), D1067–D1072 (2011)
7. D'haeseleer, P.: How does gene expression clustering work? Nat. Biotechnol. **23**(12), 1499–1501 (2005). http://www.nature.com/doifinder/10.1038/nbt1205-1499
8. Fern, X.Z., Brodley, C.E.: Random projection for high dimensional data clustering: a cluster ensemble approach. ICML **3**, 186–193 (2003)
9. Galdi, P., Napolitano, F., Tagliaferri, R.: Consensus clustering in gene expression. In: Serio, C., Liò, P., Nonis, A., Tagliaferri, R. (eds.) CIBB 2014. LNCS, vol. 8623, pp. 57–67. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24462-4_5
10. Gautier, E.L., Shay, T., Miller, J., Greter, M., Jakubzick, C., Ivanov, S., Helft, J., Chow, A., Elpek, K.G., Gordonov, S., et al.: Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages. Nat. Immunol. **13**(11), 1118–1128 (2012)
11. Handl, J., Knowles, J., Kell, D.B.: Computational cluster validation in post-genomic data analysis. Bioinformatics **21**(15), 3201–3212 (2005). (Oxford, England). http://www.ncbi.nlm.nih.gov/pubmed/15914541
12. Hecht-Nielsen, R.: Context vectors: general purpose approximate meaning representations self-organized from raw data. In: Computational Intelligence: Imitating Life, pp. 43–56 (1994)

13. Johnson, W.B., Lindenstrauss, J.: Extensions of lipschitz mappings into a hilbert space. Contemp. Math. **26**(189–206), 1 (1984)
14. Kimes, P.K., Cabanski, C.R., Wilkerson, M.D., Zhao, N., Johnson, A.R., Perou, C.M., Makowski, L., Maher, C.A., Liu, Y., Marron, J.S., et al.: SigFuge: single gene clustering of RNA-seq reveals differential isoform usage among cancer samples. Nucleic Acids Res. **42**(14), e113–e113 (2014)
15. Kuncheva, L.I., Hadjitodorov, S.T.: Using diversity in cluster ensembles. In: 2004 IEEE International Conference on Systems, Man and Cybernetics, vol. 2, pp. 1214–1219. IEEE (2004)
16. Lam, Y.K., Tsang, P.W.: eXploratory K-Means: a new simple and efficient algorithm for gene clustering. Appl. Soft Comput. **12**(3), 1149–1157 (2012)
17. Monti, S., Tamayo, P., Mesirov, J., Golub, T.: Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach. Learn. **52**(1/2), 91–118 (2003). http://link.springer.com/10.1023/A:1023949509487
18. Rodriguez, J., Kuncheva, L., Alonso, C.: Rotation forest: a new classifier ensemble method. IEEE Trans. Pattern Anal. Mach. Intell. **28**(10), 1619–1630 (2006). http://ieeexplore.ieee.org/document/1677518/
19. Serra, A., Fratello, M., Fortino, V., Raiconi, G., Tagliaferri, R., Greco, D.: MVDA: a multi-view genomic data integration methodology. BMC Bioinform. **16**(1), 1 (2015)
20. Shen, R., Mo, Q., Schultz, N., Seshan, V.E., Olshen, A.B., Huse, J., Ladanyi, M., Sander, C.: Integrative subtype discovery in glioblastoma using icluster. PLoS ONE **7**(4), e35236 (2012)
21. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Nat. Acad. Sci. US Am. **102**(43), 15545–15550 (2005). http://www.ncbi.nlm.nih.gov/pubmed/16199517, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1239896
22. Vega-Pons, S., Ruiz-Shulcloper, J.: A survey of clustering ensemble algorithms. Int. J. Pattern Recogn. Artif. Intell. **25**(03), 337–372 (2011). http://www.worldscientific.com/doi/abs/10.1142/S0218001411008683
23. 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. Nat. Methods **11**(3), 333–337 (2014). http://www.nature.com/doifinder/10.1038/nmeth.2810