Serra *et al. BMC Bioinformatics* (2015) 16:261
DOI 10.1186/s12859-015-0680-3

BMC Bioinformatics

RESEARCH ARTICLE Open Access

MVDA: a multi-view genomic data integration methodology

Angela Serra[1], Michele Fratello[2], Vittorio Fortino[3], Giancarlo Raiconi[1], Roberto Tagliaferri[1†] and Dario Greco[3*†]

Abstract

Background: Multiple high-throughput molecular profiling by omics technologies can be collected for the same individuals. Combining these data, rather than exploiting them separately, can significantly increase the power of clinically relevant patients subclassifications.

Results: We propose a multi-view approach in which the information from different data layers (views) is integrated at the levels of the results of each single view clustering iterations. It works by factorizing the membership matrices in a late integration manner. We evaluated the effectiveness and the performance of our method on six multi-view cancer datasets. In all the cases, we found patient sub-classes with statistical significance, identifying novel sub-groups previously not emphasized in literature. Our method performed better as compared to other multi-view clustering algorithms and, unlike other existing methods, it is able to quantify the contribution of single views on the final results.

Conclusion: Our observations suggest that integration of prior information with genomic features in the subtyping analysis is an effective strategy in identifying disease subgroups. The methodology is implemented in R and the source code is available online at http://neuronelab.unisa.it/a-multi-view-genomic-data-integration-methodology/.

Keywords: Clustering, Multi-view, Subclasses

Background

Stratifying patients into distinct subgroups can lead to more accurate diagnostic and treatment strategies. Current methods for patient stratification are usually based on gene expression data and apply cluster algorithms to identify groups of patients having similar expression profiles [1–3]. For example, multivariate gene expression signatures have been shown to discriminate between disease subtypes, such as recurrent and non-recurrent cancer types or tumour progression stages [4]. In addition to gene expression data other omics data types, such as miRNA (microRNA) expression, methylation or copy number alterations, can be used to improve the model accuracy for patient stratification. For example, somatic copy number alterations provide good biomarkers for cancer subtype classification [5]. Data integration approaches to efficiently identify subtypes among existing samples has recently gained attention. The main idea is to identify groups of samples that share relevant molecular characteristics. Strategies of data integration of multiple omics data types poses several computational challenges, as they deal with data having generally a small number of samples and different pre-processing strategies for each data source. Moreover, they have to cope with redundant data as well as the retrieval of the most relevant information contained in the different data sources.

Methods for clustering multiple data layers can be grouped into three main categories, namely early, intermediate, and late integration. Early integration methods directly combine all features into a single dataset [6–8]; intermediate integration methods build joint representations of data given the views [9]; late integration methods preprocess separately each individual view, subsequently combining the results [10, 11]. Late integration methods are often preferred when combining continuous and discrete data together, such as CNV and mRNA. Omics data are highly dimensional data and subject to non-Gaussian noise. Therefore, integrating them with an early or intermediate integration techniques may lead to highly noisy

*Correspondence: dario.greco@ttl.fi
†Equal contributor
[3]Unit of Systems Toxicology and Nanosafety Research Centre, Finnish Institute of Occupational Health, FIOH, Helsinki, Finland
Full list of author information is available at the end of the article

patterns unless appropriate regularization techniques are used which, however, lead to a very complex multi-view learning process.

A number of data integration approaches for patients subgroups discovery were recently proposed, based on supervised classification, unsupervised clustering or biclustering. These methodologies are called multi-view learning [12]. Examples of supervised approaches are [13, 14]. Multi-view biclustering has been used in a cocaine user subtyping [15]. Finally multi-view clustering methodologies have been intensively used also if in few cases on omics data. Multi-view clustering applied to biological data includes iCluster [16] and SNF [9]. iCluster uses a joint latent-variable model to identify the grouping structure in multi omics data. On the other hand, SNF uses a network-based approach to combine different omics data (e.g., mRNA expression, DNA methylation and microRNA expression data) to identify relevant patient subtypes. However, the contribution of the individual data sources to the classification output is not quantified in any of these multi-view clustering methods.

In this study, we propose a new computational framework for multi-view clustering that aims to combine dimensional reduction, variable selection, clustering (for each available data type) and data integration methods to find patient subtypes, as described in Fig. 1.

First, the cluster-based correlation analysis is used to reduce the number of features for each data type (genes, miRNAs, protein, etc.). Second, a ranked-based method is employed to select the features based on their ability to separate patient subtypes. Third, clustering is used to identify patient subtypes independently from each

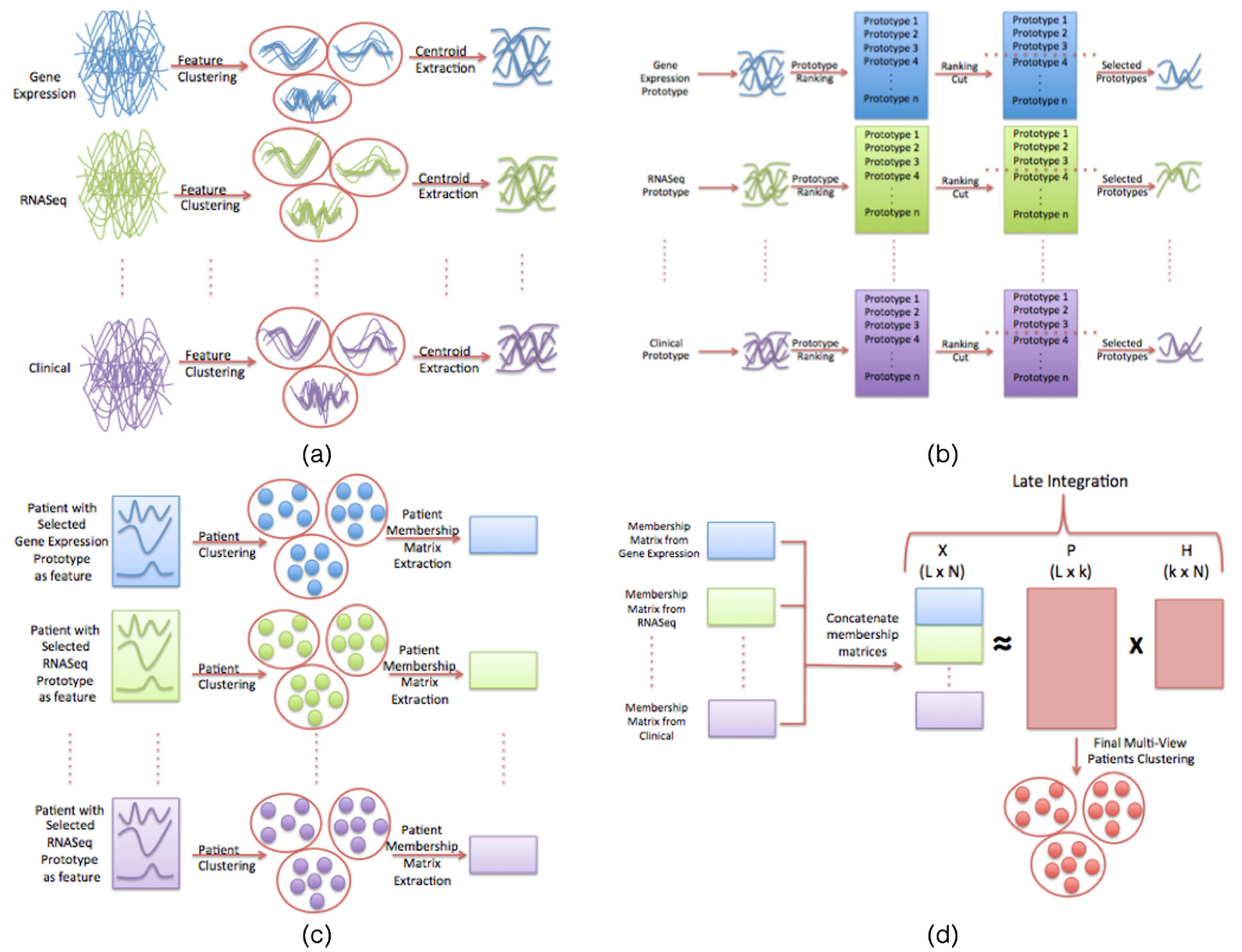

Gene Expression
Feature Clustering
Centroid Extraction
RNASeq
Feature Clustering
Centroid Extraction
Clinical
Feature Clustering
Centroid Extraction
(a)
Gene Expression Prototype
Prototype Ranking
Prototype 1
Prototype 2
Prototype 3
Prototype 4
Prototype n
Ranking Cut
Selected Prototypes
RNASeq Prototype
Clinical Prototype
(b)
Patient with Selected Gene Expression Prototype as feature
Patient Clustering
Patient Membership Matrix Extraction
Patient with Selected RNASeq Prototype as feature
Patient with Selected RNASeq Prototype as feature
(c)
Late Integration
Membership Matrix from Gene Expression
Membership Matrix from RNASeq
Membership Matrix from Clinical
Concatenate membership matrices
X
(L x N)
P
(L x k)
H
(k x N)
≈
x
Final Multi-View Patients Clustering
(d)

Fig. 1 The proposed approach: The computational approach is composed of four steps. First, the data is pre-processed. In each view feature with low variance are filtered out. Furthermore, the features are clustered in order to reduce the input dimension. From each cluster prototype are extracted. These prototypes are the only features used in following steps (**a**). Second, the prototypes are ranked by the patient class separability and the most significant ones are selected (**b**). Third, the patients are clustered and the membership matrices are obtained (**c**). Fourth, a late integration approach is utilized to integrate clustering results (**d**)

reduced dataset. Fourth, integrative clustering methods are exploited to find more robust patient subtypes and assess the contributions of different data types used for the identification of all the patient subtypes. Detailed information on each step can be found in Additional files 1 and 2. We tested our method on large genomic data sets including different omics data types, such as the Cancer Genome Atlas (TCGA) data sets (http://cancergenome.nih.gov/). Our comparison experiments suggest that our method outperforms other existing integration methods, such as Tw-Kmeans [7] and SNF [9].

Results and Discussion

We developed a novel methodology for cluster analysis of multiple genomic data types.

We compared it with recently developed methods: the integrative clustering algorithm, namely SNF [9]. and the Tw-Kmeans [7], an early integration multi-view clustering model. Using TCGA datasets from 4 different tumor types (Table 1), we evaluated the cluster impurity error, the Normalized Mutual Information [17] and the cluster stability of all the considered algorithms.

The evaluation metrics computed for each dataset are summarized in Table 3. Our unsupervised method shows a mean error of 27,47 %, normalized mutual information (NMI) of 28 % and stability of 85 %. Moreover, the error can significantly decrease when using prior information. Indeed, our method with prior information reduces the error to 6,30 %. The other methods used in the comparison study show a higher mean error from the lowest 30,83 % of SNF to the highest 30,93 % of Kmeans. They also show a lower NMI (the maximum value reached is 26 % of Ward's method) and variable stability from the lowest 51 % of the Kmeans to the highest 96 % of the partitioning around medoids (pamk).

The class label and the p-value for each cluster obtained after the integrative step is reported in Fig. 2, where the label indicates the subclass to which patients in the cluster belong, while the p-value measures the statistical significance of a cluster. In the case of the dataset OXF.BRC.1, the patients are divided into four classes: LumA, LumB, Her2 and Basal. We observed eight relevant clusters, four of which are subclasses of class LumA (cluster 4 - pvalue 2.50×10^{-4}; cluster 5 - pvalue 8.71×10^{-8}; cluster 6 - pvalue $2:92 \times 10^{-3}$; cluster 11 - pvalue 1.97×10^{-3}) and two are subclasses of class LumB (cluster 2 - pvalue $3:93 \times 10^{-14}$; cluster 10 - pvalue $5:14 \times 10^{-3}$). We also report the influence of each data on the final cluster. While it is obvious that the clusters are obtained considering all the genomic data views, the information needed to identify a specific subclass can be more relevant in a particular data type instead over the others. For example, the clusters 3, 6 and 11 of the OXF.BRC.1 dataset are both labeled as LumA. miRNA expression contributes for the 100 % to define the cluster 11, the gene expression is mainly determining the cluster 3 (57 %), while for cluster 6 they are equally important. This could mean, for example, that patients in cluster 11 are particularly characterized by miRNA expression while patients in cluster 3 by gene expression.

As shown in Fig. 3, the integrative clustering performed generally better that the clustering on each single data view. In the TCGA.BRCA dataset, the mean cluster impurity is about 26 % when patients are grouped by the gene expression and 43 % when they are grouped by their miRNA expression profiles. However, combining the gene and the miRNA expression profiles, 26,50 % of error in unsupervised mode and 9 % in semi-supervised mode are obtained, respectively. Only in a few cases, the patient grouping based on a single data view performs better than the one obtained with multiple data types.

Figure 4 depicts the comparison between the two integration methods, either with or without prior information. The matrix factorization based method reaches the higher stability (about 85 %) in all the cases. With respect to the cluster impurity, the difference is almost always negligible. The greatest difference occurs when passing from the unsupervised to the semi-supervised approach. The cluster impurity for the unsupervised clustering is about 30 % and about 7 % for semi-supervised. Therefore, for more accurate sub-typing of classes semi-supervised integration was used, which maintains high stability and reduces the classification error compared to the classes. However, in case of unbalanced patient classes, the prior information is needed to increase the prediction.

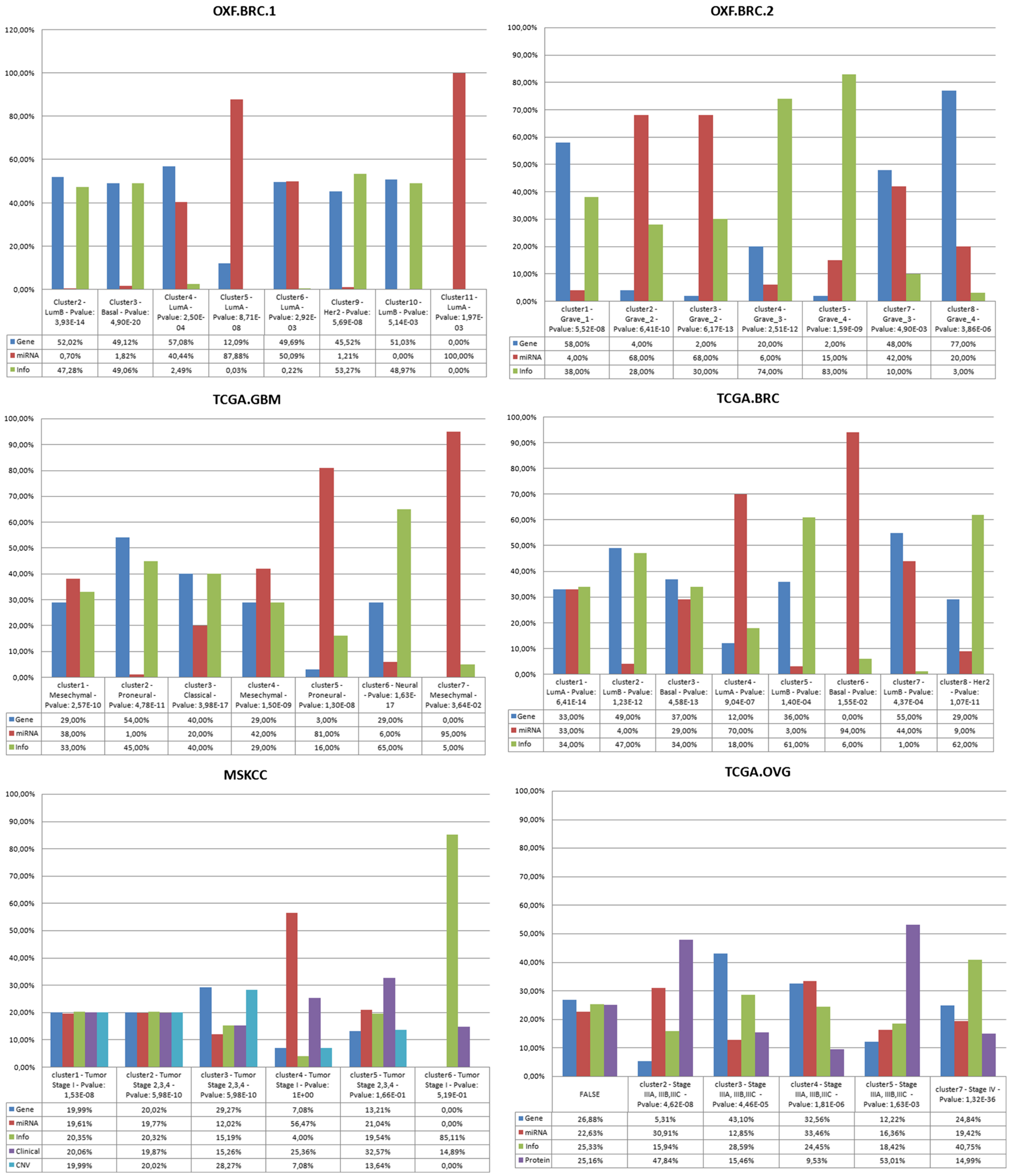
Since we tested different algorithms at each step of our methodology, we aimed at understanding if a common pipeline for all the datasets could be applied. After the execution of all the analyses, we observed that the best algorithms for the first and second steps strongly depend on the data. We found that K-means is the best algorithm for step 3 for the TCGA.BRACA, OXF.BRCA.1 and OXF.BRCA.2 datasets (Table 2). At the last step, the matrix factorization approach provided lower errors and greater stability as compared with the general linear integration methods on the majority of the datasets. This result corroborates our hypothesis that a late integration approach is better for it allows using the best algorithms for each data type.

In order to evaluate the performance of the proposed method, we systematically compared it with Tw-Kmeans and SNF algorithms (Table 3). Anyhow, we did not compare our method with iClust, as it has been show to have worse performance than SNF, with which we deal in this study [9]. We confirmed that late integration works more efficiently in integrating different views of genomic data. This is due to the large complexity and difference

Table 1 Datasets: Description of the datasets used in this study

Dataset	Response	N(0)	N(1)	N(2)	N(3)	Gene expression	RNASeq	microRNA expression	miRNASeq	Protein expression	Copy number	Clinical data
Breast Cancer from The Cancer genome Atlas, N = 151												
TCGA.BRC	Pam50 (Her2,Basal,LumA,LumB)	24	13	55	59		x		x			
Breast Cancer from The Gene Expression Omnibus, N = 201												
OXF.BRC.1	Pam50 (Her2,Basal,LumA,LumB)	26	6	117	52	x		x				
OXF.BRC.2	Clinical (Level1, Level2, Level3, Level4)	73	54	42	32	x		x				
Prostate Cancer from Memorial Sloan-Kettering Cancer Center, N = 88												
MSKCC.PRCA	Tumor stages T1 vs. T2, T3, T4	53	35			x		x			x	x
Ovarian Cancer from The Cancer Genome Atlas, N = 398												
TCGA.OVG	Tumor stage I,II, Tumor stage III, Tumor stage IV	33	315	50		x		x		x		
Glioblastoma Multiforme from The Cancer genome Atlas, N = 167												
from TCGA.GBM	(Classical, Mesechymal, Neural, Proneural)	37	54	24	52	x		x				

"N" is the number of subjects for each dataset. Ni is the number of samples in the i-th class. An x denotes if that view (column) is available for a specific dataset (row)



OXF.BRC.1

	Cluster2 - LumB - Pvalue: 3,93E-14	Cluster3 - Basal - Pvalue: 4,90E-20	Cluster4 - LumA - Pvalue: 2,50E-04	Cluster5 - LumA - Pvalue: 8,71E-08	Cluster6 - LumA - Pvalue: 2,92E-03	Cluster9 - Her2 - Pvalue: 5,69E-08	Cluster10 - LumB - Pvalue: 5,14E-03	Cluster11 - LumA - Pvalue: 1,97E-03
Gene	52,02%	49,12%	57,08%	12,09%	49,69%	45,52%	51,03%	0,00%
miRNA	0,70%	1,82%	40,44%	87,88%	50,09%	1,21%	0,00%	100,00%
Info	47,28%	49,06%	2,49%	0,03%	0,22%	53,27%	48,97%	0,00%

OXF.BRC.2

	cluster1 - Grave_1 - Pvalue: 5,52E-08	cluster2 - Grave_2 - Pvalue: 6,41E-10	cluster3 - Grave_2 - Pvalue: 6,17E-13	cluster4 - Grave_3 - Pvalue: 2,51E-12	cluster5 - Grave_4 - Pvalue: 1,59E-09	cluster7 - Grave_3 - Pvalue: 4,90E-03	cluster8 - Grave_4 - Pvalue: 3,86E-06
Gene	58,00%	4,00%	2,00%	20,00%	2,00%	48,00%	77,00%
miRNA	4,00%	68,00%	68,00%	6,00%	15,00%	42,00%	20,00%
Info	38,00%	28,00%	30,00%	74,00%	83,00%	10,00%	3,00%

TCGA.GBM

	cluster1 - Mesechymal - Pvalue: 2,57E-10	cluster2 - Proneural - Pvalue: 4,78E-11	cluster3 - Classical - Pvalue: 3,98E-17	cluster4 - Mesechymal - Pvalue: 1,50E-09	cluster5 - Proneural - Pvalue: 1,30E-08	cluster6 - Neural - Pvalue: 1,63E-17	cluster7 - Mesechymal - Pvalue: 3,64E-02
Gene	29,00%	54,00%	40,00%	29,00%	3,00%	29,00%	0,00%
miRNA	38,00%	1,00%	20,00%	42,00%	81,00%	6,00%	95,00%
Info	33,00%	45,00%	40,00%	29,00%	16,00%	65,00%	5,00%

TCGA.BRC

	cluster1 - LumA - Pvalue: 6,41E-14	cluster2 - LumB - Pvalue: 1,23E-12	cluster3 - Basal - Pvalue: 4,58E-13	cluster4 - LumA - Pvalue: 9,04E-07	cluster5 - LumB - Pvalue: 1,40E-04	cluster6 - Basal - Pvalue: 1,55E-02	cluster7 - LumB - Pvalue: 4,37E-04	cluster8 - Her2 - Pvalue: 1,07E-11
Gene	33,00%	49,00%	37,00%	12,00%	36,00%	0,00%	55,00%	29,00%
miRNA	33,00%	4,00%	29,00%	70,00%	3,00%	94,00%	44,00%	9,00%
Info	34,00%	47,00%	34,00%	18,00%	61,00%	6,00%	1,00%	62,00%

MSKCC

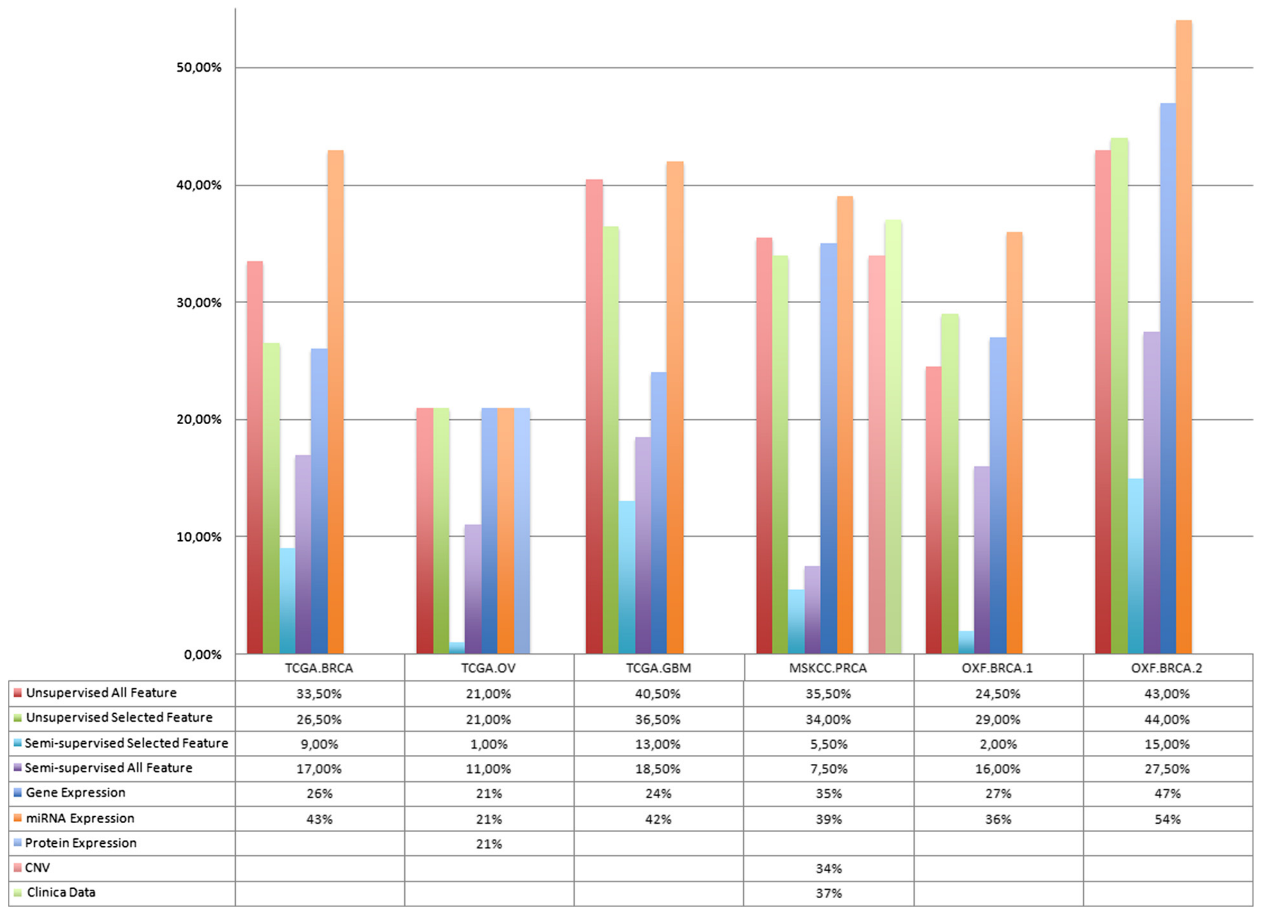
	cluster1 - Tumor Stage I - Pvalue: 1,53E-08	cluster2 - Tumor Stage 2,3,4 - Pvalue: 5,98E-10	cluster3 - Tumor Stage 2,3,4 - Pvalue: 5,98E-10	cluster4 - Tumor Stage I - Pvalue: 1E+00	cluster5 - Tumor Stage 2,3,4 - Pvalue: 1,66E-01	cluster6 - Tumor Stage I - Pvalue: 5,19E-01
Gene	19,99%	20,02%	29,27%	7,08%	13,21%	0,00%
miRNA	19,61%	19,77%	12,02%	56,47%	21,04%	0,00%
Info	20,35%	20,32%	15,19%	4,00%	19,54%	85,11%
Clinical	20,06%	19,87%	15,26%	25,36%	32,57%	14,89%
CNV	19,99%	20,02%	28,27%	7,08%	13,64%	0,00%

TCGA.OVG

	FALSE	cluster2 - Stage IIIA, IIIB,IIIC - Pvalue: 4,62E-08	cluster3 - Stage IIIA, IIIB,IIIC - Pvalue: 4,46E-05	cluster4 - Stage IIIA, IIIB,IIIC - Pvalue: 1,81E-06	cluster5 - Stage IIIA, IIIB,IIIC - Pvalue: 1,63E-03	cluster7 - Stage IV - Pvalue: 1,32E-36
Gene	26,88%	5,31%	43,10%	32,56%	12,22%	24,84%
miRNA	22,63%	30,91%	12,85%	33,46%	16,36%	19,42%
Info	25,33%	15,94%	28,59%	24,45%	18,42%	40,75%
Protein	25,16%	47,84%	15,46%	9,53%	53,01%	14,99%

Fig. 2 Multi-View Clusters Statistics: For each cluster class label, the p-value and the view contribution are reported. For all the six datasets, the results showed that the matrix factorization method gives lower classification error and better accuracy than the approach with general linear integration

between the views. When views have different numerical and statistical characterizations, it is more convenient to individually analyze single data types and then combine the results in a multi-view analysis. This becomes more and more important as the number of views involved in the analysis increases.



	TCGA.BRCA	TCGA.OV	TCGA.GBM	MSKCC.PRCA	OXF.BRCA.1	OXF.BRCA.2
■ Unsupervised All Feature	33,50%	21,00%	40,50%	35,50%	24,50%	43,00%
■ Unsupervised Selected Feature	26,50%	21,00%	36,50%	34,00%	29,00%	44,00%
■ Semi-supervised Selected Feature	9,00%	1,00%	13,00%	5,50%	2,00%	15,00%
■ Semi-supervised All Feature	17,00%	11,00%	18,50%	7,50%	16,00%	27,50%
■ Gene Expression	26%	21%	24%	35%	27%	47%
■ miRNA Expression	43%	21%	42%	39%	36%	54%
■ Protein Expression		21%				
■ CNV				34%		
■ Clinica Data				37%		

Fig. 3 Cluster Impurity difference between single view and integration analysis: Cluster impurity was evaluated as the fraction of objects that were inconsistent with the label of the cluster. It was calculated using each data type alone and by integrating them. Errors decreased with the integration approach in particular when the semi-supervised methodologies were used

Evaluation of genes in breast cancer datasets

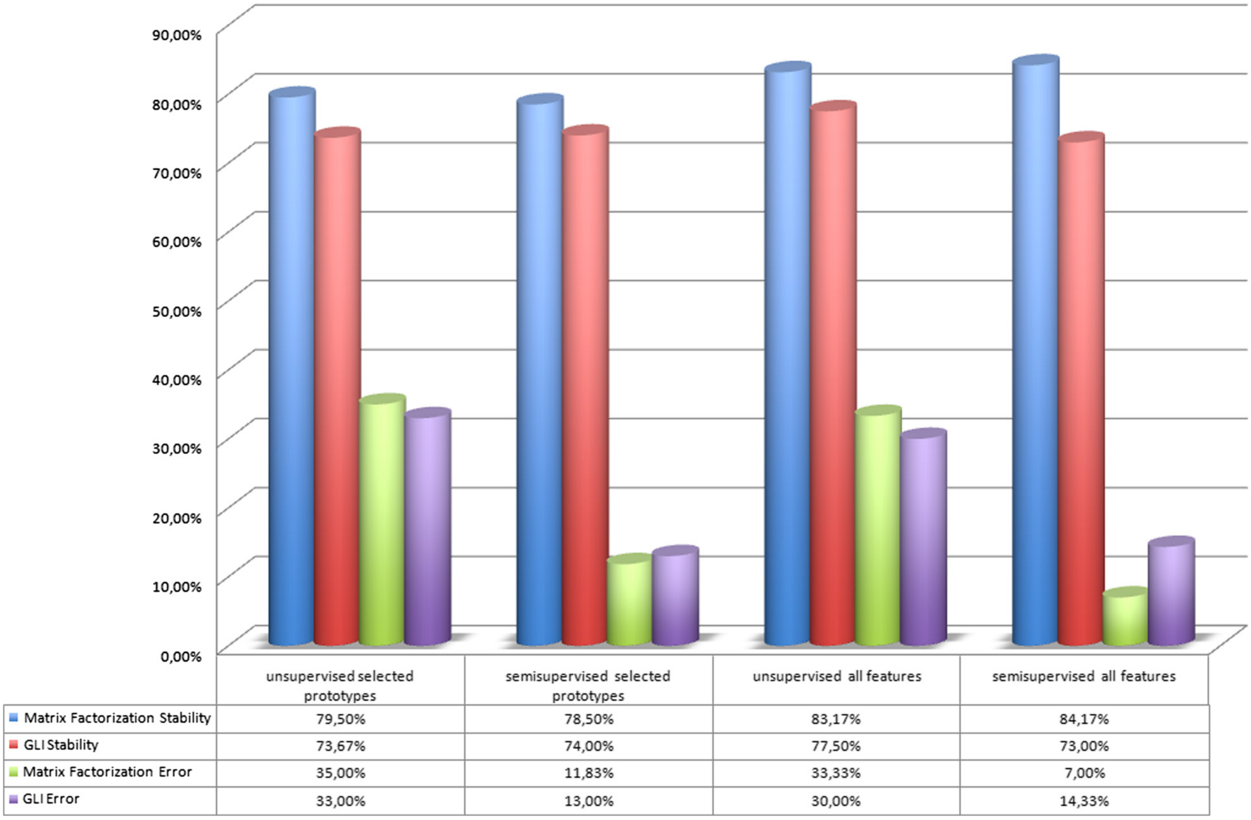
We selected a robust set of features from each analyzed dataset in order to find common features (Fig. 5a) and highlight shared patterns by enrichment analysis (Fig. 5b). Each list of features was obtained by using the Borda-count rule across the leave-one-out replicates. The enrichment analysis was performed by using the DAVID functional annotation tool [18, 19] and graphically displayed with the R package BACA [20]. Figure 5b reports a chart indicating unique and common Gene Ontology (GO) terms found by using DAVID on the different lists. It is possible observe that the three lists of features highlight similar GO annotations, involved for instance in regulation of kinase activity and regulation of cellcycle. The list of genes shared between the three breast cancer datasets can be found in Additional file 3.

Conclusions

In this study, we proposed a methodology for multiple genomic data type analysis aiming at patients subtyping. The methodology is composed of four steps using state of the art algorithms. Furthermore we systematically searched for the best algorithm for each step on six of benchmark datasets. We performed experiments in a late integration fashion, with two different algorithms. Since we were interested in high accuracy in class patient subtyping, we used prior information as a new view in the integration process. We found that the integrative clustering outperforms the single view approaches on all the datasets. We also showed that our method is stable by executing clustering on perturbed datasets removing one patient at a time and evaluating the normalized mutual information between all the resulting clusterings.

Methods

The proposed methodology for the analysis of multi-view biological datasets takes in input n matrices $M_i \in R^{F_i \times P}$ *for* $i = 1, \ldots, n$, where F_i is the number of features (genes, miRNAs, CNV, methylation, clinical information, etc.) and P is the number of patients and a vector cl



	unsupervised selected prototypes	semisupervised selected prototypes	unsupervised all features	semisupervised all features
🟦 Matrix Factorization Stability	79,50%	78,50%	83,17%	84,17%
🟥 GLI Stability	73,67%	74,00%	77,50%	73,00%
🟩 Matrix Factorization Error	35,00%	11,83%	33,33%	7,00%
🟪 GLI Error	33,00%	13,00%	30,00%	14,33%

Fig. 4 Difference between alternative integration methods: The mean cluster stability is reported, as calculated on four covariates represented by the type of experiment executed. Clustering stability was calculated by comparing the unsupervised and the semi-supervised mode, both using either all the features or only the selected prototypes

Table 2 Best combination of methods for each step: Summary of the best combination of algorithms for each view used to obtain the best grouping of patients that identifies significant sub-classes

Dataset	Views	(a) Feature clustering	(b) Feature selection	(c) Patients clustering	(d) Late integration
TCGA.BRCA	RNASeq	Pam	CAT-score	Kmeans	MF
	miRNASeq	Pam	CAT-score	Pam	
TCGA.OV	Gene Expression	Pam	Random Forest	DM	MF
	Protein Expression	Pam	-	DM	
	miRNA Expression	Pam	-	DM	
TCGA.GBM	Gene Expressions	Spectral	CAT-score	Kmeans	MF
	miRNA Expression	Ward	-	Kmeans	
OXF.BRCA.1	Gene Expressions	Pam	Random Forest	Ward	GLI
	miRNA Expression	Pam	Random Forest	Kmeans	
OXF.BRCA.2	Gene Expressions	Pvcluster	CAT-score	Kmeans	MF
	miRNA Expressions	Pam	Random Forest	Kmeans	
MSKCC	Gene Expressions	Pam	CAT-score	Kmeans	MF
	miRNA Expressions	Pam	-	Pam	
	CNV	Spectral	CAT-score	Kmeans	
	Clinical	-	-	Pam	

In the feature selection column the symbol (-) means that feature selection was not executed because the number of features was small. Symbol (DM) in Patient clustering column means that same classification error was obtained with all the algorithms used

Table 3 Validation Results: The mean classification error, normalized mutual information (NMI) and stability, on all datasets, are shown, measuring the agreement between the clusters resulting from an approach and the real patient classification

	Feature	Integration	Algorithm	Error	NMI	Stability
Single View	All Feature	-	Ward	30,08 %	26 %	86 %
		-	Kmeans	30,93 %	25 %	51 %
		-	Pamk	30,75 %	24 %	94 %
	Selected Prototype	-	Ward	30,72 %	26 %	89 %
		-	Kmeans	30,36 %	25 %	52 %
		-	Pamk	30,78 %	24 %	96 %
Multi-View	All Feature	Early Integration	Tw-kmeans	37,10 %	24 %	69 %
	All Feature	Intermediate Integration	SNF	30,83 %	22 %	83 %
	All Feature in Cluster of Selected Prototype	Intermediate Integration	SNF	31,31 %	18 %	82 %
	Selected Prototype	Late Integration unsupervised	MF/GLI	**27,47 %**	**28 %**	**85 %**
	Selected Prototype	Late Integration semi-supervised	MF/GLI	**6,30 %**	**63 %**	**84 %**

Bold font in percentage indicates best performance in the experiments of classes labels, and yields a multi-view partitioning $G = \bigcup_{i=1}^{k}(G_i)$ of patients. The multi-view integration methods also return a matrix C where $c[i,j]$ is the contribution of view i to the final multi-view cluster j.

The approach consists of four main steps as shown in Fig. 1:

1. Prototype Extraction: for each view, the features were filtered by variance and clustered in order to find prototypes, reducing the input data dimension.
2. Prototype ranking: the prototypes found in the step 1 were ranked based on their ability to separate the classes.
3. Single view clustering: in each view, the samples were clustered using the prototypes created in the steps 1 and 2 as features
4. Integration: single view clustering results were integrated with a late integration approach, in order to obtain the k final multi-view meta-clusters

The late integration methodology can be considered as a further step of the proposed data mining pipeline, in which the clustering results of each single view are unified. This approach offers a number of significant advantages: (i) clustering algorithms can be optimally chosen with respect to each single view; (ii) it can be naturally parallelized; (iii) representation issues are avoided since clustering results are the inputs to the integration algorithms.

Prototype extraction

The features with low variance across the samples were eliminated. Therefore the data were clustered with respect to the patients and the cluster centroids were selected as the prototype patterns. The centroid of each cluster was selected as the most correlated element with respect to the other elements in that cluster. Different clustering algorithms were used: Pvclust [21], SOM [22], hierarchical clustering with Ward's method [23], K-means [24], Partitional Around Medoids [25] and Spectral clustering [26].

The idea is to evaluate several popular clustering techniques and compare their behaviour on the different views with respect to the hierarchical method that is the standard algorithm used to cluster genes. As noted in [27], cluster analysis is a complex and interactive process and results change based on its parameters. Therefore, each algorithm was executed for different values of K. For each algorithm and for each K, clustering performance was evaluated according to the following evaluation function:

$$VAL = \frac{1}{4}\left(\frac{IC+1}{2} + 1 - \frac{EC+1}{2} + (1-S) + CG\right) \quad (1)$$

where IC is the complete diameter measure, representing the average sample correlation of the less similar objects in the same cluster; EC is the complete linkage measure, representing the average sample correlation of the less similar objects for each pair of clusters; S is the singleton factor and CG is the compression gain. The evaluation function was defined in order to obtain the output value normalized between 0 and 1. The complete diameter and the complete linkage measures were calculated with the R "clv" package [28]. The number of singleton was normalized in a range $(0, 1)$ in order to be comparable with the correlation measure. It was defined as $S = N/(K-1)$. The compression gain was defined as $CG = 1 - (K/N_{elem})$, where K is the number of clusters and N is the number of elements to be clustered.

Each clustering algorithm was executed on n different values of K and the corresponding results were evaluated with the function VAL. Values close to 1 indicate a clustering with similar objects in the clusters, weakly linked clusters, with few singletons and with a good compression rate. A numeric score was then assigned to each K value by considering the average values of the VAL function compiled over the clustering results obtained with the different algorithms. Then, the K showing the highest score was chosen and subsequently used to identify the best clustering algorithms having the first two highest scores with respect to the selected k value. In Algorithm 1 is reported the computational procedure followed to fine-tuned the k-values for the cluster analysis.

Algorithm 1 Procedure for the k-values fine-tuning

```
procedure CHOOSEK(k $\in K = (k_1, k_2, \cdots, k_n)$, data)
OUTPUT: BestK: the optimal number of clusters;
    BestAlgo1, BestAlgo2: The first two best performing algorithms;
    for all k $\in K = (k_1, k_2, \cdots, k_n)$ do
        $L_{pv}^{k} \leftarrow \mathrm{VAL}(Pvclust(data, k))$
        $L_{km}^{k} \leftarrow \mathrm{VAL}(Kmeans(data, k))$
        $L_{s}^{k} \leftarrow \mathrm{VAL}(SOM(data, k))$
        $L_{sp}^{k} \leftarrow \mathrm{VAL}(Spectral(data, k))$
        $L_{p}^{k} \leftarrow \mathrm{VAL}(Pam(data, k))$
        $L_{w}^{k} \leftarrow \mathrm{VAL}(Ward(data, k))$
        $S^{k} \leftarrow \mathrm{MEAN}\left(L_{pv}^{k}, L_{km}^{k}, L_{s}^{k}, L_{sp}^{k}, L_{p}^{k}, L_{w}^{k}\right)$
    end for
    $BestK \leftarrow \mathrm{WHICH.MAX}(S^{k})$
    $BestAlgo_1 \leftarrow \mathrm{WHICH.MAX}\left(\left\{L_{pv}^{BesK}, L_{km}^{BesK}, L_{s}^{BesK}, L_{sp}^{BesK}, L_{p}^{BesK}, L_{w}^{BesK}\right\}\right)$
    $BestAlgo_2 \leftarrow \mathrm{WHICH.MAX}\left(\left\{L_{pv}^{BesK}, L_{km}^{BesK}, L_{s}^{BesK}, L_{sp}^{BesK}, L_{p}^{BesK}, L_{w}^{BesK}\right\} \setminus \left\{BestAlgo_1\right\}\right)$
end procedure
```

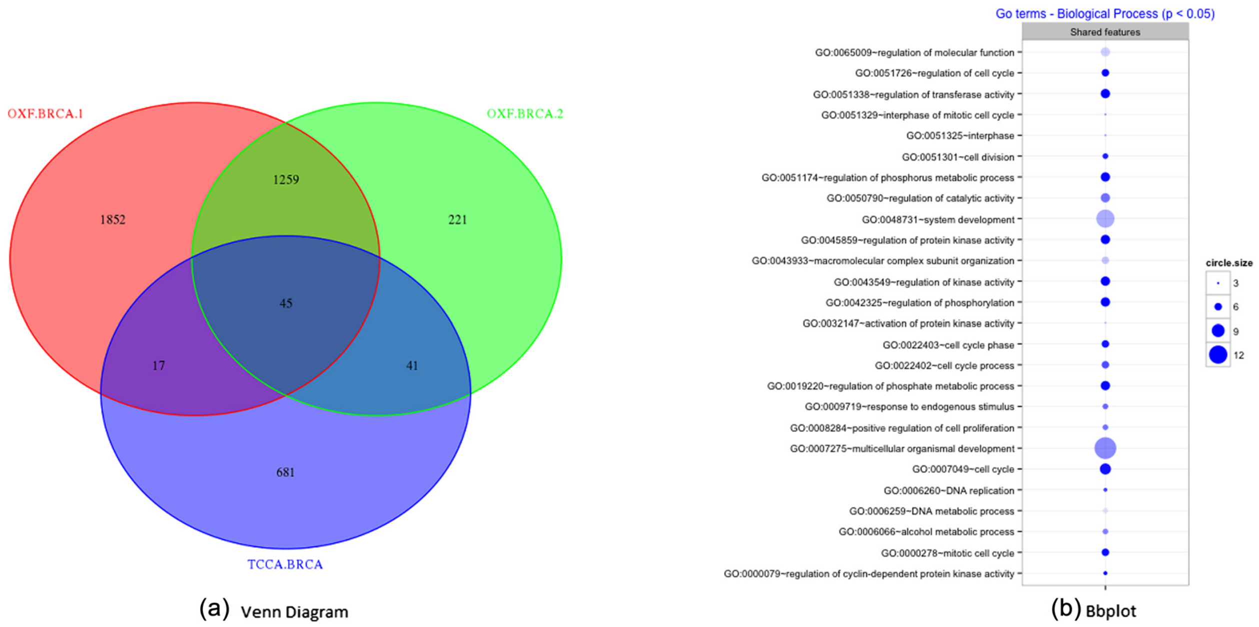


Fig. 5 Breast Cancer Gene Analysis: (**a**) the Venn diagram shows the number of common relevant genes between the three datasets. The analysis highlights 45 common genes between the three lists. (**b**) The bubble plot displays the enriched GO terms found by using DAVID. A transparent bubble indicates a set of significant genes, a dark bubble indicates a set of highly significant genes. The diameter of the bubble indicates the number of genes related to the same GO term

Feature ranking

If the number of prototypes, after the fist step, was still high, further dimensional reduction by feature selection was done. Feature ranking was performed by computing the CAT-score [29] and the Mean Decreasing Accuracy index calculated by Random Forests [30]. The parameters of RF-based classifiers were fine-tuned by using the R package rminer [31]. It provides a function that first tunes the hyper parameter(s) of a selected model by using bootstrap methods and subsequently builds the corresponding supervised data-mining model. For each rank, the cumulative sum of the ranking score was computed and four different cuts based on the cumulative values were taken. Cuts took into account all the features needed to maintain 60 %, 70 %, 80 % and 90 % of the cumulative value. An example is shown in section Prototype Extraction of Additional file 1. These different groups of features were used to cluster patients in each single view, with the same single view clustering algorithms used in the previous step. The number of clusters K was considered as the number of classes. For each clustering, the error was calculated as the dispersion obtained in the confusion matrix between class labels and clustering assignments. The clustering algorithm that reached the minimum error for each view was then selected. These clustering results were used as the input to the late integration step.

Integration

Two late integration methods were used: the matrix factorization approach [11] and a general model for multiview integration [10]. The first method [11] combines information by factorizing the membership matrix of patient single-view clusterings. The method starts by transposing all the membership matrices and stacking them vertically obtaining the matrix of cluster $X \in R^{lXn}$ where l is the total number of cluster in C. The objective is to find the best approximation of X such that

$$X = PH \; and \; P >= 0, H >= 0 \tag{2}$$

The results of the factorization are two matrices: $P \in R^{lXk}$ that projects the clusters in a new set of k meta-clusters and $H \in R^{kXn}$ whose columns can be viewed as the membership of the original objects in the new set of meta-clusters. Based on the values in the projection matrix P, we can calculate a matrix $T \in R^{vXk}$. T_{hf} indicates the contribution of the view V_h to the f-th meta-cluster. Based on values in P it is also possible to find the optimal value of k for the number of multi-view clusters we want in output. The matrix factorization was run with a range of values for k as input and the algorithm returns the factorization for the best value of k.

The second method exploits the intuition that the optimal clustering is the consensus clustering shared by as many views as possible. This can be reformulated as an optimization problem where the optimal clustering is the closest to all the single view clusterings under a certain distance or dissimilarity measure. Clusterings are again represented as membership matrices.

Formally the model can be described as follow: given a set of clustering membership matrices $M = [M_1, \ldots, M_h] \in R_+^{n \times l}$ and a positive integer k, the optimal clustering membership matrix $B \in R_+^{n \times k}$ and the optimal mapping matrices $P = [P_1, \ldots, P_h] \in R_+^{k \times l}$ are given by the minimization:

$$\begin{aligned} \min_{B,P} \quad & GI(M||BP) \\ \text{s. t.} \quad & P \geq 0 \\ & B \geq 0, B\mathbf{1} = \mathbf{1} \end{aligned} \tag{3}$$

where $GI(M||BP)$ is the generalized Kullback-Leibler divergence such that

$$GI(X||Y) = \sum_{ij} \left(logX_{ij}log\frac{X_{ij}}{Y_{ij}} - X_{ij} + Y_{ij} \right)$$

subject to the constraint that both P and B must be non-negative and that each row of B must sum to one.

By taking the membership matrix for each of the previous clusterings, and, using these two late integration methods, a multi-view clustering was obtained. Experiments were performed in two ways: the former uses all the prototypes for classification; the latter uses only the most relevant ones for class separability. Each one of these approaches were performed both in unsupervised and semi-supervised manners, respectively.

The semi-supervised approach consists of giving a priori information as input to the techniques of late integration via a membership matrix of patients with the exact information of their classes. This information is combined with the membership of the patients compared to the single view clustering and integrated in metaclusters. This can be a useful approach mainly when the data set is composed of unbalanced or under represented classes.

Derivation of subclasses

Once the multi-view clusters were obtained, a subclass was assigned to each one. For each cluster, the number of objects of each class was calculated and the class with more representative patterns was assigned as the cluster label. Then, a p-value was calculated in order to verify the statistical significance of the subclass by the Fisher's exact test [32].

Validation

The method was compared with classical single view clustering algorithms, early and intermediate integration approaches. For each method clustering impurity, normalized mutual information (NMI) and cluster stability were evaluated. Cluster impurity was defined as the number of patients in the cluster whose label differs from that of the cluster. Given two clustering solutions Cl_1 and Cl_2 NMI was computed as the mutual information between the two clustering normalized by the cluster entropies. The NMI was computed between clustering results and real patient classifications.

Since prior information was introduced, the stability of the system was tested with leave-one-out technique. A test in itself was run on the first step to generate a stability index for the prototypes of the obtained clusters. Then, the steps 2, 3 and 4 were evaluated jointly to assess the stability of the selected features and to evaluate the robustness of the multi-view clustering results. Furthermore, a borda-count [33] method was performed to find the final list of features selected over the leave-one-out experiments for the integration step.

At the end of this process, N different clustering assignments were obtained, one for each removed patient. An $N \times N$ matrix M was created, where $M(i,j)$ was the normalized mutual information (NMI) between the clustering obtained removing patient i and the clustering obtained removing patient *j*. Then the mean of the matrix M was calculated, indicating the stability measure of the method.

The comparison study involved the following methods:

- Kmeans, Hierarchical and Pam single view clustering
- Tw-Kmeans, an early integration multi-view clustering algorithm
- SNF, an intermediate integration multi-view clustering algorithm

Experiments with single view clustering algorithms were executed in feature concatenation mode: data from views were concatenated and used as a new greater feature space. This kind of experiments were run both on the most variable features for each view and on the most relevant prototypes found after the first and second steps of our approach. Experiments with Tw-kmeans were executed on all the features without any manipulation of the initial datasets. Experiments with SNF were executed both using all the features and using all the features that belong to the clusters associated to the relevant prototypes.

Dataset collection and preparation

Six datasets were downloaded from The Cancer Genome Atlas (TCGA) (https://tcga-data.nci.nih.gov/tcga/), Memoral Sloan-Kettering Cancer Center (http://cbio.mskcc.org/) and from NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/) (See Table 1).

TCGA.BRC

Breast cancer dataset from the TCGA repository (https://tcga-data.nci.nih.gov/tcga/ - Breast invasive carcinoma [BRCA]). The samples in this dataset correspond to breast cancer patients with invasive tumors. Genomic data for two views were downloaded: RNASeq and miRNASeq (Level 3). Because level 3 data corresponds to already preprocessed data, only the batch effect was removed by the comBat method in the R "sva" package [34]. Patients were subsequently divided into four classes (Her2, Basal, LumA, LumB), using PAM50 classifier [35, 36].

OXF.BRC.1

Breast cancer dataset from a study performed at Oxford University [37]. Data were downloaded from Gene Expression Omnibus Dataset (http://www.ncbi.nlm.nih.gov/geo/). Data were available for two views: mRNA and microRNA expression under the accession number GSE22219 and GSE22220. Patients were divided into four classes (Her2, Basal, LumA, LumB), using PAM50 classifier [35, 36].

OXF.BRC.2

Breast cancer dataset from a study performed at Oxford University [37]. Data were downloaded from Gene Expression Omnibus Dataset (http://www.ncbi.nlm.nih.gov/geo/). Data were available for two views: mRNA and microRNA expression under the accession number GSE22219 and GSE22220. Patients were divided into four classes (Level1, Level2, Level3, Level4) using clinical data also retrieved from the same source. See Table 4 for classes definition.

TCGA.GBM

Glioblastoma cancer dataset from the TCGA repository. The samples in this dataset correspond to glioblastoma patient with invasive tumors. TCGA website was accessed (https://tcga-data.nci.nih.gov/tcga/ - Glioblastoma multiforme [GBM]) and publicly available data for two views were downloaded: gene expression and miRNA expression. Also clinical data was retrieved. The patients were divided info four classes: Classical, Mesechymal, Neural and Proneural as described in [38].

Table 4 Oxford Dataset: Oxford Dataset, class definition by clinical data

Class	Clinical information
Level1	er = 1, node = 0, grade = 1–2
	er = 1, node = 0, grade = 3–4
Level2	er = 1, node > 0, grade = 1–2
	er = 1, node > 0, grade = 3–4
Level3	er = 0, node = 0, grade = 1–2
	er = 0, node = 0, grade = 3–4
Level4	er = 0, node > 0, grade = 1–2
	er = 0, node > 0, grade = 3–4

TCGA.OVG

Ovarian cancer dataset from the TCGA repository (https://tcga-data.nci.nih.gov/tcga/ - Ovarian serous cystadenocarcinoma [OV]). The samples in this dataset correspond to patient affected by ovarian serous cystadenocarcinoma tumors. Publicly available data for three views were downloaded: gene expression, protein expression, and miRNA expression. Clinical data were downloaded in order to classify patients in three categories. In particular patients were classified by clinical stage: first class: stage IA, IB, IC, IIA, IIB and IIC, second class: IIIA, IIIB and IIIC, third class Stage IV.

MSKCC.PRCA

Prostate cancer dataset from a study performed at the Memorial Sloan Kettering Cancer Center (http://cbio.mskcc.org/). The samples in these datasets correspond to patient prostate cancer tumors. The MSKCC Cancer Genomics data portal (http://cbio.mskcc.org/cancergenomics/prostate/data/) was accessed and data for five views were downloaded: clinical data, gene expression, microRNA expression and copy number variation. Patients were classified in two classes by using clinical data by the tumor stage: class one is Tumor Stage I and class two is Tumor Stage II, III and IV. Classification of patient was done according to a previous study performed on the same dataset [14].

Additional files

Additional file 1: It contains a section for each step of the methodology in which the tables and figures with the results for each dataset are reported. (PDF 1495 kb)

Additional file 2: Each sheet refers to each dataset analysed, reporting the results of the single-view clustering patients. Clustering errors for each algorithm and each cut of feature are also reported. (XLSX 54 kb)

Additional file 3: It contains the gene symbols and description for all shared genes between the tree breast cancer datasets highlighted by the analysis. (DOCX 14 kb)

Abbreviations

miRNA: MicroRNA; CNV: Copy number variation; NMI: Normalized mutual information; GO: Gene Ontology; SOM: Self-organizing map; Pam: Partitional around medoids; CAT-score: Correlation-adjusted t-score; DAVID: Database for annotation, visualization and integrated discovery; TCGA: The Cancer Genome Atlas; MSKCC: Memorial Sloan Kettering Cancer Center; GEO: Gene Expression Omnibus.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DG and RT conceived and supervised the study. AS developed the methods, analysed and interpreted the data, and implemented the software. MF, VF and

GR participated the development of the methods and the analysis of the data. All the authors have participated in drafting the manuscript. All authors read and approved the final manuscript.

Acknowledgements
This work has been supported by the European Commission, under grant agreement FP7-309329 (NANOSOLUTIONS).
The results shown here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

Author details
[1]NeuRoNe Lab, Department of Computer Science, University of Salerno, Fisciano, Italy. [2]Department of Medical, Surgical, Neurological, Metabolic and Ageing Sciences, Second University of Napoli, Napoli, Italy. [3]Unit of Systems Toxicology and Nanosafety Research Centre, Finnish Institute of Occupational Health, FIOH, Helsinki, Finland.

Received: 2 February 2015 Accepted: 20 July 2015
Published online: 19 August 2015

References

1. Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, Sørlie T, et al. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc National Acad Sci U S A. 2005;102(10):3738–43.
2. Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, et al. Gene expression predictors of breast cancer outcomes. The Lancet. 2003;361(9369):1590–6.
3. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc National Acad Sci. 2001;98(20):11462–7.
4. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc National Acad Sci. 2001;98(19):10869–74.
5. Vang Nielsen K, Ejlertsen B, Møller S, Trøst Jørgensen J, Knoop A, Knudsen H, et al. The value of top2a gene copy number variation as a biomarker in breast cancer: Update of dbcg trial 89d. Acta Oncologica. 2008;47(4):725–34.
6. Kailing K, Kriegel HP, Pryakhin A, Schubert M. Clustering multi-represented objects with noise. In: Advances in Knowledge Discovery and Data Mining. Berlin Heidelberg: Springer; 2004. p. 394–403.
7. Chen X, Xu X, Huang JZ, Ye Y. Tw-(k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013;25(4):932–44.
8. Sa VRD. Spectral Clustering with Two Views. In: ICML workshop on learning with multiple views; 2005. p. 20–27.
9. Wang B, Mezlini AM, 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. 2014;11(3):333–7.
10. Long B, Yu PS, Zhang Z. A general model for multiple view unsupervised learning. In: Society for Industrial and Applied Mathematics - 8th SIAM International Conference on Data Mining 2008, Proceedings in Applied Mathematics; 2008. p. 822–33.
11. Greene D. A Matrix Factorization Approach for Integrating Multiple Data Views. Mach Learn Knowl Discov Databases. 2009;5781:423–38.
12. Xu C, Tao D, Xu C. A survey on multi-view learning. 2013. arXiv preprint arXiv:1304.5634.
13. Wasito I, Istiqlal A, Budi I. Data integration model for cancer subtype identification using Kernel Dimensionality Reduction-Support Vector Machine (KDR-SVM). In: Computing and Convergence Technology (ICCCT), 2012 7th International Conference On. IEEE; 2012. p. 876–80.
14. Ray B, Henaff M, Ma S, Efstathiadis E, Peskin ER, Picone M, et al. Information content and analysis methods for multi-modal high-throughput biomedical data. Sci Rep. 2014;4:4411. doi:10.1038/srep04411.
15. Sun J, Bi J, Kranzler HR. Multi-view singular value decomposition for disease subtyping and genetic associations. BMC Genet. 2014;15(1):73.
16. Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J, et al. Integrative subtype discovery in glioblastoma using icluster. PLoS ONE. 2012;7(4): 35236.
17. Vinh NX, Epps J, Bailey J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. J Mach Learn Res. 2010;11:2837–854.
18. Dennis Jr G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. David: database for annotation, visualization, and integrated discovery. Genome Biol. 2003;4(5):3.
19. Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, et al. David bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 2007;35(suppl 2):169–75.
20. Fortino V, Alenius H, Greco D. Baca: bubble chart to compare annotations. BMC Bioinformatics. 2015;16(1):37.
21. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics (Oxford England). 2006;22(12):1540–2. doi:10.1093/bioinformatics/btl117.
22. Vesanto J, Alhoniemi E. Clustering of the self-organizing map. Neural Netw IEEE Trans. 2000;11(3):586–600.
23. Ward JH. Hierarchical Grouping to Optimize an Objective Function. J Am Stat Assoc. 1963;58(301):236–44. doi:10.1080/01621459.1963.10500845.
24. Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. J R Stat Soc. 1979;28:100–8. doi:10.2307/2346830.
25. Kaufman L, Rousseeuw PJ. Clustering by means of medoids. In: Data analysis based on the L 1-Norm and related methods. North-Holland; 1987. p. 405–416.
26. Ng AY, Jordan MI, Weiss Y. On spectral clustering: Analysis and an algorithm. Adv Neural Inf Process Systs. 2002;2:849–56.
27. Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. Bioinformatics (Oxford, England). 2005;21(15):3201–12. doi:10.1093/bioinformatics/bti517.
28. Nieweglowski L. Clv: Cluster Validation Techniques. 2013. R package version 0.3-2.1. http://CRAN.R-project.org/package=clv.
29. Ahdesmäki M, Strimmer K. Feature selection in omics prediction problems using cat scores and false nondiscovery rate control. Ann Appl Stat. 2010;4(1):503–19. doi:10.1214/09-AOAS277. arXiv:0903.2003v4.
30. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
31. Cortez P. Rminer: Data Mining Classification and Regression Methods. 2014. R package version 1.4. http://CRAN.R-project.org/package=rminer.
32. Fisher RA. JSTOR: J R Stat Soc. 1922;85(1):87–94. http://www.jstor.org/stable/2340521?__redirected. Accessed 17/06/14.
33. Lin S. Space oriented rank-based data integration. Stat Appl Genet Mol Biol. 2010;9(1):1544-6115. doi: 10.2202/1544-6115.1534, April 2010.
34. Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Storey JD. Sva: Surrogate Variable Analysis. R package version 3.14.0.
35. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A. 2002;99(10):6567–72. doi:10.1073/pnas.082099299.
36. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol Off J Am Soc Clin Oncol. 2009;27(8):1160–7. doi:10.1200/JCO.2008.18.1370.
37. Buffa FM, Camps C, Winchester L, Snell CE, Gee HE, Sheldon H, et al. microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. Cancer Res. 2011;71(17):5635–45. doi:10.1158/0008-5472.CAN-11-0489.
38. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010;17(1):98–110. doi:10.1016/j.ccr.2009.12.020.