EL-Manzalawy et al. BMC Medical Genomics 2018, 11(Suppl 3):71
https://doi.org/10.1186/s12920-018-0388-0

BMC Medical Genomics

RESEARCH Open Access

Min-redundancy and max-relevance multi-view feature selection for predicting ovarian cancer survival using multi-omics data

Yasser EL-Manzalawy[1,5,6], Tsung-Yu Hsieh[1,4,5], Manu Shivakumar[2], Dokyoon Kim[2,3*] and Vasant Honavar[1,3,4,5,6*]

From The 7th Translational Bioinformatics Conference
Los Angeles, CA, USA. 29 September - 01 October 2017

Abstract

Background: Large-scale collaborative precision medicine initiatives (e.g., The Cancer Genome Atlas (TCGA)) are yielding rich multi-omics data. Integrative analyses of the resulting multi-omics data, such as somatic mutation, copy number alteration (CNA), DNA methylation, miRNA, gene expression, and protein expression, offer tantalizing possibilities for realizing the promise and potential of precision medicine in cancer prevention, diagnosis, and treatment by substantially improving our understanding of underlying mechanisms as well as the discovery of novel biomarkers for different types of cancers. However, such analyses present a number of challenges, including heterogeneity, and high-dimensionality of omics data.

Methods: We propose a novel framework for multi-omics data integration using multi-view feature selection. We introduce a novel multi-view feature selection algorithm, MRMR-mv, an adaptation of the well-known Min-Redundancy and Maximum-Relevance (MRMR) single-view feature selection algorithm to the multi-view setting.

Results: We report results of experiments using an ovarian cancer multi-omics dataset derived from the TCGA database on the task of predicting ovarian cancer survival. Our results suggest that multi-view models outperform both view-specific models (i.e., models trained and tested using a single type of omics data) and models based on two baseline data fusion methods.

Conclusions: Our results demonstrate the potential of multi-view feature selection in integrative analyses and predictive modeling from multi-omics data.

Keywords: Multi-omics data integration, Multi-view feature selection, Cancer survival prediction, Machine learning

* Correspondence: dkim@geisinger.edu; vhonavar@ist.psu.edu
[2]Biomedical and Translational Informatics Institute, Geisinger Health System, Danville, PA, USA
[1]Artificial Intelligence Research Laboratory, College of Information Sciences and Technology, Pennsylvania State University, University Park, PA 16802, USA
Full list of author information is available at the end of the article

Background

The advent of "big data" offers enormous potential for understanding and predicting health risks and intervention outcomes, as well as personalizing treatments, through integrative analysis of clinical, biomedical, behavioral, environmental, and even socio-demographic data. For example, recent efforts in cancer genomics under the Precision Health Initiative offer promising ways to diagnose, prevent, and treat many cancers [1]. Recent advances in high-throughput omics technologies offer cost-effective ways to acquire diverse types of genome-wide multi-omics data. For instance, Large-scale collaborative efforts such as the Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) are collecting multi-omics data for tumors along with clinical data for the patients. An important goal of these initiatives is to develop comprehensive catalogs of key genomic alterations associated for a large number of cancer types [2, 3].

Computational analyses of multi-omics data offer an unprecedented opportunity to deepen our understanding of complex underlying mechanisms of cancer that is essential for advancing precision oncology (See for example, [4–7]). Because different types of omics data have been shown to complement each other [8], there is a growing interest in effective methods for integrative analyses of multi-omics data [9–11]. The resulting methods have been successfully used to predict the molecular abnormalities that impact both clinical outcomes and therapeutic targets [5, 10, 12–16].

Effective approaches to integrative analyses and predictive modeling from multi-omics data have to address three major challenges [5]: i) the curse of dimensionality (i.e., the number of features p is very large compared to the number of samples n); ii) the differences in scales as well as sampling/collection bias and noise present in different omics data sets; iii) extracting and optimally combining, for the prediction task at hand, features that provide complementary information across different data sources. Unfortunately, baseline methods that simply concatenate the features extracted from the different data sources or analyze each data from each source separately and combine the predictions fail to satisfactorily address these challenges. Therefore, there is an urgent need for more sophisticated methods for integrative analysis and predictive modeling from multi-omics data [16].

The problem of learning predictive models from multi-omics data can be naturally formulated as a *multi-view learning* problem [17] where each omics data source provides a distinct view of the complex biological system. Multi-view learning offers a promising approach to developing predictive models by leveraging complementary information provided by the multiple data sources (views) to optimize the predictive performance of the resulting model [17]. The state-of-the-art learning algorithms attempt to learn a set of models, one from each view, and combine them so as to jointly optimize the predictive performance of the combined multi-view model. Some examples of multi-view learning algorithms include: multi-view support vector machines [18], multi-view Boosting [19], multi-view k-means [20], and clustering via canonical correlation analysis [21]. However, barring a few exceptions (e.g., multi-view feature selection methods [22], and multi-view representation learning [23]) the vast majority of existing multi-view learning algorithms are not equipped to effectively cope with the high-dimensionality of omics data [17]. Hence, predictive modeling from multi-omics data calls for effective methods for multi-view feature selection or dimensionality reduction.

Against this background, we present a general two-stage framework for multi-omics data integration. We introduce MRMR-mv, an adaptation of the well-known Min-Redundancy and Maximum-Relevance (MRMR) single-view feature selection algorithm to the multi-view setting. We provide, to the best of our knowledge, the first application of a multi-view feature selection method to predictive modeling from multi-omics data. We report the results of our experiments that compare the proposed approach with several baseline methods on the task of building a predictive model of cancer survival [13] using a TCGA multi-omics dataset composed of three omics data sources, copy number alteration (CNA), DNA methylation, and gene expression RNA-Seq. The results of our experiments show that: (i) the multi-view predictive models developed from multi-omics data outperform their single-view counterparts; and that (ii) the predictive models developed using MRMR-mv for multi-view feature selection outperform those developed using two baseline methods that combine multiple views into a single-view. These results demonstrate the potential of multi-view feature selection based approaches to multi-omics data integration.

Methods

Datasets

Normalized and preprocessed multi-omics ovarian cancer datasets (most recently updated on August 16, 2016), including genelevel copy number alteration (CNA), DNA methylation, and gene expression (GE) RNA-Seq data, were downloaded from UCSC Xena cancer genomic browser [24]. Table 1 summarizes the number of samples and features (e.g., genes) in each dataset. Clinical data about vital status and survival for the subjects were also downloaded from Xena server. Only the patients with CNA, methylation, RNA-Seq, and survival data were retained. Patients with survival time ≥3 years were labeled as long-term survivors while patients with survival time <3 years and vital status of 0 were labeled as short-term

Table 1 TCGA ovarian cancer omics data used in this study

Data source	Platform	Number of samples	Number of features	Number of features with high variance
CNA	Affymetrix SNP 6	579	24,777	7355
Methylation	Illumina Infinium HumanMethylation27k	616	27,579	6206
GE RNA-Seq	Illumina HiSeq	308	30,531	283

survivors. The resulting multi-view dataset consists of 215 samples, 127 of them are classified as long-term survivors. Each view was then pre-filtered and normalized as follows: i) features with missing values were excluded; ii) feature values in each sample were rescaled to lie in the interval [0,1]; iii) features with variance less than 0.02 were removed.

Notations

Table 2 summarizes convenient notations used in this work. For simplicity, we assumed a binary label for each sample. Note however, that Algorithms 1 and 2, described below, are also applicable to multi-class as well as numerically labeled data.

Minimum redundancy and maximum relevance feature selection

Unlike univariate feature selection methods [25] that return a subset of features without accounting for redundancy between the selected features, the minimum redundancy and maximum relevance (MRMR) feature selection algorithm [26] iteratively selects features that are *maximally relevant* for the prediction task and *minimally redundant* with the set of already selected features. MRMR has been successfully used for feature selection in a number of applications including microarray gene expression data analysis [26, 27], prediction of protein sub-cellular localization [28], epileptic seizure [29], and protein-protein interaction [30].

While the exact solution to the problem of MRMR selection of $k = |S|$ features from a set of n candidates requires the evaluation of $O(n^k)$ candidate feature subsets, it is possible to obtain an approximate solution using a simple heuristic algorithm (see Algorithm 1) [26]. Algorithm 1 accepts as input: a labeled dataset D; a function $g:(x_i, x_j) \rightarrow R^+$ that quantifies the redundancy between any pair of features (e.g., the absolute value of Pearson's correlation coefficient); a function $f:(x_i, y) \rightarrow R^+$ that quantifies the relevance of a target feature for predicting the labels y (e.g., mutual information (MI) or F-statistic); and the number of features k to be selected using the MRMR criterion. In lines 1 and 2, the algorithm creates an empty set S and the feature with the maximum relevance for predicting y is added to S. In each of the subsequent $k - 1$ iterations (lines 3–5), the features that greedily approximate the MRMR criterion in Eq. 1 are successively

Table 2 Notations

Symbol	Definition and Description
$D = < X, y>$	Labeled dataset where $X \in R^{m \times n}$ is a matrix of m instances and n features, and $y \in \{0, 1\}^m$ is the binary class labels of the instances
x_i	i^{th} feature in X
$g(x_i, x_j)$	Function that returns the redundancy between two features x_i and x_j
$f(x_i, y)$	Function that returns the relevance between a feature x_i and class labels y
S	Indices of selected features
Ω	Indices of all features
Ω_S	Indices of candidate features $\Omega - S$
k	Number of features to be selected
v	Number of views in a multi-view dataset
$MVD = < (X^1, \ldots, X^v), y>$	Labeled multi-view dataset where $X^i \in R^{m \times n_i}$ is a matrix of m samples and n_i features and $y \in \{0, 1\}^m$ is the binary class labels of the instances in all views
$D^i = < X^i, y>$	i^{th} view in a multi-view dataset
x^i_j	j^{th} feature in X^i
S^i	Indices of selected features from i^{th} view
Ω^i	Indices of all features in i^{th} view
Ω_{S^i}	Indices of candidate features $\Omega^i - S^i$ in i^{th} view

added to S. Eq. 1 has two terms: the first term maximizes the relevance condition, whereas the second term minimizes the redundancy condition.

$$\text{argmax}_{j\in\Omega_S}\left(f(x_j,y)-\frac{1}{|S|^2}\sum_{l\in S}g(x_j,x_l)\right) \quad (1)$$

Algorithm 1. MRMR

Require: $D =< X, y >, g, f, k$

1: $S \leftarrow \emptyset$

2: add $x_i = \text{argmax}_{j\in\Omega} f(x_j, y)$ to S

3: **for** $t = 1 : k-1$ do

4: add the feature that satisfies Eq. 1 to S

5: **end for**

6: **return** S

Multi-view minimum redundancy and maximum relevance feature selection

MRMR, or any single-view feature selection algorithm, can be trivially applied to multi-view data as follows: i) Apply MRMR separately to each view and then concatenate view-specific selected features. The major limitation of this approach is that it ignores the redundancy and complementarity of features across the views [31]; ii) Apply MRMR to a single-view dataset obtained by concatenating all the views. A key limitation of this approach is that it fails to explicitly account for the prediction task specific differences in the relative utility or relevance of the features extracted from the different views.

Here, we propose a novel multi-view feature selection algorithm, MRMR-mv, that adapts the MRMR algorithm to the multi-view setting. MRMR-mv (shown in Algorithm 2) accepts as input: a labeled multi-view dataset, *MVD*, with $v \geq 2$ views; a redundancy function g; a relevance function f; number of features to be selected k; and a probability distribution $P = \{p_1 \cdots p_v\}$ that models the relative importance of each view (or the prior probability that a view contributes a feature to the set of features selected by MRMR-mv). If each of the views is equally important, P should be a uniform distribution. MRMR-mv proceeds as follows. First, S^t is initialized for each view t to keep track of selected features from that view (lines 1–3). Second, the procedure *choice*, implemented in NumPy python library [32], is invoked to obtain k-1 views, sampled from with replacement, according to P, from the set of views. The list of sampled views is recorded in C (lines 4 and 5). Third, the maximally relevant feature across *all* of the views (say x_j^i, the j^{th} feature in the i^{th} view) is retrieved and the set (S^i) of the selected features for the corresponding view, i, is updated accordingly (line 6). Fourth, for each of the views in C, considered in turn and at each step t, the feature from the corresponding view that satisfies the MRMR criterion with respect to the previously selected features from iterations (1 through t-1) is added to $S^{C[t]}$ (lines 7–10). Finally, the algorithm returns selected view-specific features $S^1, \cdots S^v$.

We note that MRMR-mv applies the MRMR criteria across all of the views, as opposed to the baseline methods that apply the criteria to each view separately or to the concatenation of all views. Thus MRMR-mv can select complementary features from within as well as across views. It can also assign different degrees of importance to the views to reflect any available information about their relative utility in the context of a given prediction task.

Algorithm 2. MRMR-mv

Require: $MVD =< (X^1, \ldots, X^v), y >, g, f, k, P = (p_i, \ldots, p_v)$

1: for $t = 1 : v$

2: $S^t \leftarrow \emptyset$

3: end for

4: $V \leftarrow \{1, \ldots, v\}$

5: $C \leftarrow choice(V, k-1, P)$

6: add $x_j^i = \text{argmax}_{i\in\{1,\ldots,v\}} \text{argmax}_{j\in\Omega^i} f(x_j^i, y)$ to S^i

7: **for** $t = 1 : k-1$ do

8: $l \leftarrow C[t]$

9: add $x_i^l = \text{argmax}_{j\in\Omega_{s^l}}(f(x_j^l, y) - \frac{1}{(t)^2}\sum_{q\in\{1,\ldots,v\}}\sum_{h\in S^q} g(x_j^l, x_h^q))$ to S^l

10: **end for**

11: **return** $S = \{S^1, \cdots, S^v\}$

A two-stage feature selection framework for multi-omics data integration

Figure 1 shows our proposed two-stage framework for integrating multi-omics data for virtually any prediction task (e.g., predicting cancer survival or predicting clinical outcome). The input to our framework is a labeled multi-view dataset in the form $D^i = < X^i, y>$. Stage I includes view-specific filters that can be used to encapsulate any traditional single-view feature selection method (e.g., Lasso [33] or MRMR). Each filter has a gating signal that could be used to disable that filter in which case the disabled filter passes on no data to the 2nd stage. A special view-specific filter, called AllFilter, passes *all* of the input features without performing any feature selection. Stage II has a single filter that can encapsulate either a single-view or a multi-view feature selection method. If the 2nd stage filter encapsulates a single-view feature selection method, the feature selection method will be applied to the concatenation of the Stage II input. On the other hand, if the 2nd stage filter encapsulates a multi-view feature selection method (e.g., MRMR-mv), then the multi-view feature selection method will be applied to the multi-view input of Stage II. The framework supports two modes of operations: i) training mode, where each enabled filter will be

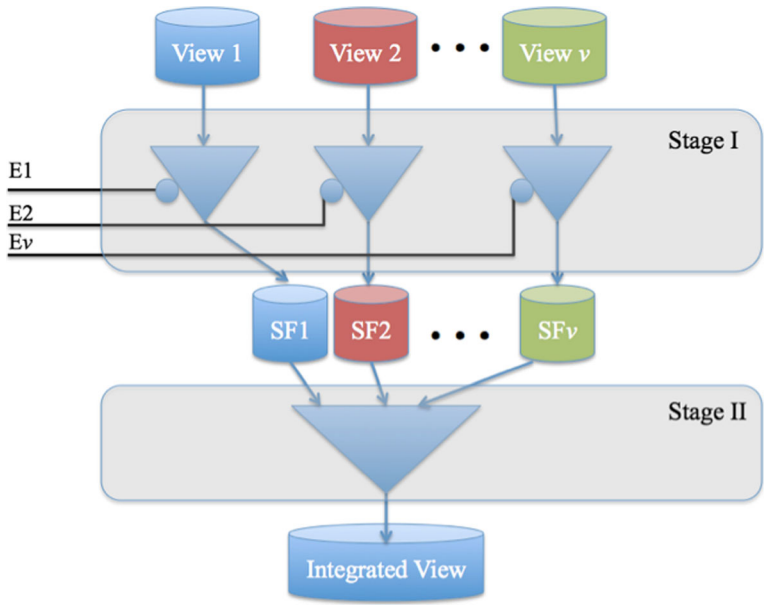

Fig. 1 Two-stage framework for integrating multi-omics data. E_i refers to the enable signal for the i^{th} view-specific filter. F_i refers to the set of features selected from the i^{th} view using the i^{th} filter

trained using the input so as to produce the filtered version of the input; ii) test (or operation) mode, where test multi-view dataset is provided as input and the trained filters will output the selected features of the input data.

The framework can be easily customized so as to allow evaluation of different approaches of predictive modeling from multi-omics data. For example, to build a single-view model by applying the Lasso method to the i^{th} view, we: set E_i to 1 and disable all other filters; pass Lasso feature selection method to the i^{th} filter; use AllFilter as Stage II filter. Similarly, to apply MRMR to concatenated views, we: enable Stage I filters and use either AllFilter (to pass the input as is) or any single-view filter; and deploy MRMR as the Stage II filter.

Implementation

We implemented Algorithms 1 and 2 and the two-stage feature selection framework in Python using the scikit-learn machine learning library [34]. We will release the code as part of sklearn-fuse, a python library for data and model-based data fusion that is currently under development in our lab. In the mean time, the code for the methods described above will be made available to interested researchers upon request.

Experiments

We report results of experiments on the task of building a predictive model of cancer survival from an ovarian cancer multi-omics dataset derived from the TCGA database. The resulting data set is comprised of three views, namely, CNA, methylation, and gene expression RNA-Seq for each patient along with the corresponding clinical outcomes (short-term versus long-term survival). Our first set of experiments consider single-view classifiers based on each of the 3 views to obtain view-specific models for comparison with the proposed multi-view models; The second set of experiments compare some of the representative instantiations of the two-stage multi-view feature selection framework in combination with some representative choices of (single-view) supervised algorithms for training the classifiers. In both cases, we experimented with three widely used machine learning algorithms for developing cancer survival predictors: i) Random Forest (RF) [35] with 500 trees; ii) eXtreme Gradient Boosting (XGB) [36] with 500 weak learners; ii) Logistic Regression (LR) [37] with L1 regularization. We used the implementations of these algorithms available in the Scikit-learn machine learning library [34].

For Stage I feature selection, we experimented with several feature selection methods implemented in Scikit-learn including: RF feature importance [35]; Lasso [33]; ElasticNet [38]; and Recursive Feature Elimination (RFE) [39]. However, due to space limitation, we describe only the results of the best performing method, Lasso with L1 regularization parameter set to 0.0001. In Stage II feature selection, we used MRMR as a baseline method and MRMR-mv for multi-view feature selection.

For both MRMR and MRMR-mv feature selection, we used the absolute value of Pearson's correlation coefficient as the redundancy function, g. For the relevance function, f, we experimented with three functions Chi2,

F-Statistic (F-Stat), and Mutual Information (MI). All functions are implemented in Scikit-learn.

We estimated the performance of the resulting classifiers on the task of predicting cancer survival using the 5-fold cross-validation (CV) procedure. Briefly, the dataset is randomly partitioned into five equal subsets. Four of the five subsets are collectively used to select the features and train the classifier and the remaining subset is held out for estimating the performance of the trained classifier. This procedure is repeated 5 times, by setting aside a different subset of the data for estimating model performance. The 5 results from all the folds are then averaged to report a single performance estimate. In our experiments we used the area under ROC curve (AUC) [40] to assess the predictive performance of classifiers. When the number of samples used to estimate the classifier performance is small, as is the case with the ovarian cancer data, the estimated performance can vary substantially across different random partitions of the data into 5 folds (see Section "Single-view models for predicting ovarian cancer survival" for details). To obtain a more robust estimate of performance, we ran the 5-fold cross-validation procedure 10 times (each using different partitioning of the data into 5 subsets) and reported the mean AUC estimated from the 10 5-fold CV experiments.

Results

Single-view models for predicting ovarian cancer survival

We evaluated RF, XGB, and LR classifiers trained using each of the individual views with the top k features selected using Lasso feature selection algorithm for choices of $k = 10,20,30, \ldots, 100$.Tables 3, 4 and 5 report the performance of the resulting classifiers averaged over 10 different 5-fold cross-validation experiments.

Table 3 Average AUC scores of RF, XGB, and LR models trained on CNA data, estimated using 10 runs of 5-fold cross validation

# Features	RF	XGB	LR
10	0.57	0.56	0.58
20	0.61	0.61	0.61
30	0.61	0.61	0.61
40	0.63	0.62	0.61
50	0.64	0.64	0.62
60	0.65	0.65	0.63
70	0.65	0.65	0.63
80	0.65	0.65	0.62
90	0.66	0.66	0.63
100	0.66	0.66	0.62
Max	0.66	0.66	0.63
Avg.	0.63	0.63	0.62

Table 4 Average AUC scores of RF, XGB, and LR models trained on methylation data, estimated using 10 runs of 5-fold cross validation

# Features	RF	XGB	LR
10	0.52	0.51	0.50
20	0.51	0.52	0.50
30	0.52	0.52	0.49
40	0.52	0.53	0.50
50	0.52	0.53	0.51
60	0.52	0.53	0.52
70	0.53	0.54	0.51
80	0.53	0.54	0.52
90	0.53	0.55	0.52
100	0.53	0.55	0.52
Max	0.53	0.55	0.52
Avg.	0.52	0.53	0.51

We observed that models built using only the methylation view performed marginally better than random guessing (i.e., the best observed average AUC in Table 5 is 0.55). In contrast, single-view models using CNA or RNA-Seq achieved higher average AUC scores of up to 0.66. These results are in agreement with those of previously reported studies (e.g., [13]). It should be noted that when the performance of single-view models is estimated using a single 5-fold cross-validation experiment (as opposed to average over 10 different cross-validation experiments), the best observed AUC scores were 0.70, 0.55, and 0.69 for models built from the CNA, methylation, and RNA-Seq views, respectively. The observed variability in performance among different 5-fold cross-validation experiments is expected because of the relatively small size of the ovarian cancer survival dataset. This finding underscores

Table 5 Average AUC scores of RF, XGB, and LR models trained on RNA-Seq data, estimated using 10 runs of 5-fold cross validation

# Features	RF	XGB	LR
10	0.58	0.57	0.59
20	0.60	0.58	0.61
30	0.61	0.60	0.63
40	0.62	0.61	0.64
50	0.62	0.61	0.65
60	0.63	0.60	0.66
70	0.63	0.60	0.64
80	0.64	0.60	0.65
90	0.63	0.61	0.65
100	0.64	0.61	0.65
Max	0.64	0.61	0.66
Avg.	0.62	0.60	0.64

the importance of using multiple CV experiments to obtain robust estimates and comparisons of classifier performance. Next, we show how integrating data sources (i.e., views) can further improve the predictive performance of the cancer survival predictors.

Integrative analyses of multi-omics data sources using multi-view feature selection

We used our two-stage feature selection framework (See Fig. 1) to construct multi-view (MV) models with the following settings. The input to the framework is two views, CNA and RNA-Seq. We chose not to use the methylation view because the performance of single-view models built using the methylation data performed marginally better than chance (see Section "Single-view models for predicting ovarian cancer survival"). For the Stage I filters, we used Lasso with L1 regularization parameter set to 0.0001 to select the top 100 features from CNA and RNA-Seq views, respectively. For the Stage II filter, we used MRMR-mv with Pearson's correlation coefficient as the redundancy function and a uniform distribution for the selection probability parameter, P. Finally, we experimented with different multi-view models obtained using combinations of choices for the remaining MRMR-mv parameters, k and f. Specifically, we experimented with $k = 10, 20, \ldots, 100$ and the relevance function $f \in \{Chi2, F-Stat, MI, and\ CFM\}$, where CFM is the average of the other three relevance functions.

Figure 2 compares the performance of the different MV models described above. Interestingly, no single relevance function consistently outperforms other functions for different choices of the number of selected features, k, and machine learning algorithms. However, the best AUC of 0.7 is obtained using either $Chi2$ or MI relevance functions and RF classifier trained using the top 100 features. Hence, our final MV models will use $Chi2$ as the relevance function and the remaining MRMR-mv settings stated in the preceding paragraph.

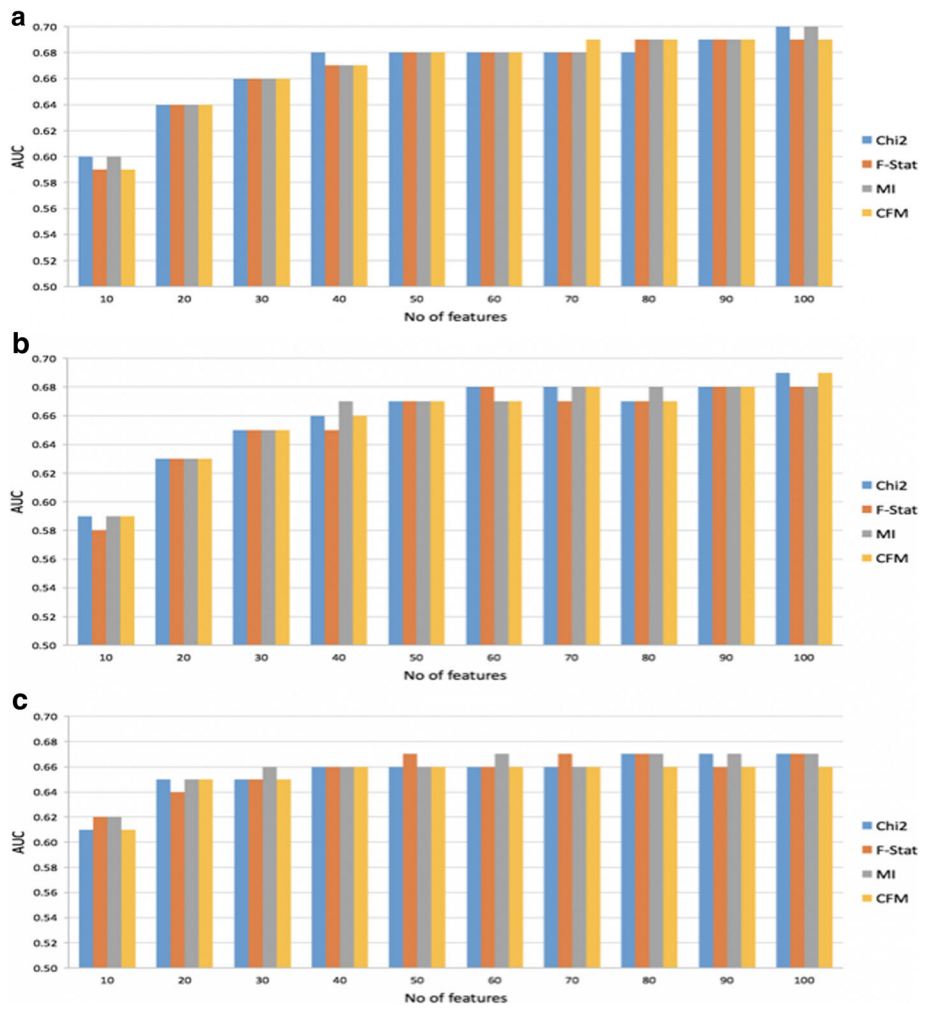

Fig. 2 Performance comparisons of multi-view models using four different relevance functions for MRMR-mv and three machine learning classifiers, **a**) RF, **b**). XGB, and **c**) LR

The selection probability parameter, *P*, in MRMR-mv algorithm controls the expected number of selected features from each view. Results shown in Fig. 2 have been produced using a uniform selection probability distribution. Although using a uniform distribution is reasonable since the best AUC score for the single-view models based on CNA or RNA-Seq is 0.66 (See Tables 3 and 5), it is interesting to examine the influence of *P* on the performance of our MV models. Let $P = (p_1, p_2)$ be the probability distribution where p_1 and p_2 denotes the sampling probability for CNA and RNA-Seq, respectively. In this experiment, we considered 11 different probability distributions obtained using $p_1 = \{0, 0.1, 0.2, \ldots, 1\}$. Then, for each choice of the number of selected features, *k*, we evaluated 11 MV models using RF algorithm and the same MRMR-mv settings described in the preceding subsection and the 11 different probability distributions for *P*. We used the percent relative range in the recorded AUC to assess the sensitivity of MV models to changes in *P*. Figure 3 shows the relationship between the number of selected MV features, *k*, and the sensitivity of MV models to changes in *P*. Interestingly, our results suggest that as the number of selected MV features increases, the resulting MV models become less sensitive to the selection probability distribution parameter *P*.

Multi-view vs. single-view models for predicting ovarian cancer survival

Figure 4 compares our final MV models with the following single-view models: i) SV_CNA, single-view models developed using CNA data source; ii) SV_RNA-Seq, single-view models developed using RNA-Seq data source; iii) SV_C, single-view models obtained by applying MRMR to the *concatenation* of the two views, CNA and RNA-Seq; iv) SV_S, single-view models obtained by applying MRMR *separately* to CNA and RNA-Seq views, respectively. In addition, Fig. 4 shows the results for a simple ensemble model that averages the predictions from SV_CNA and MV models. In general, MV and Ensemble models outperform SV models in most of the cases.

We noted some interesting observations from our experiments with each of the machine learning algorithms considered in our experiments. In the case of models developed using RF algorithm, MV and Ensemble models outperformed the four single-view models for all choices of the number of selected features, k. Ensemble models outperformed MV models for k = 10, 20, and 80. Baseline single-view models outperformed SV_CNA and SV_RNA-Seq for $k \leq 40$. The highest observed AUC was 0.7 and was obtained using the MV model and k=100. In the case of XGB based models, SV_S, MV, and Ensemble models outperformed the remaining single-view models. Ensemble models outperformed MV models for 8 out of 10 choices of k. Finally, for models developed using LR algorithm, SV_S, MV, and Ensemble models outperformed the other three single-view models. Regardless of which machine learning algorithm was used, SV_RNA--Seq and SV_C models had the lowest AUC in most of the cases reported in Fig. 4. Our results suggest that the best single-view model is more likely to perform better than models developed using concatenated views. Our results also suggest that either applying feature selection to each individual view or selecting features jointly using multi-view feature selection consistently outperform the best single-view model.

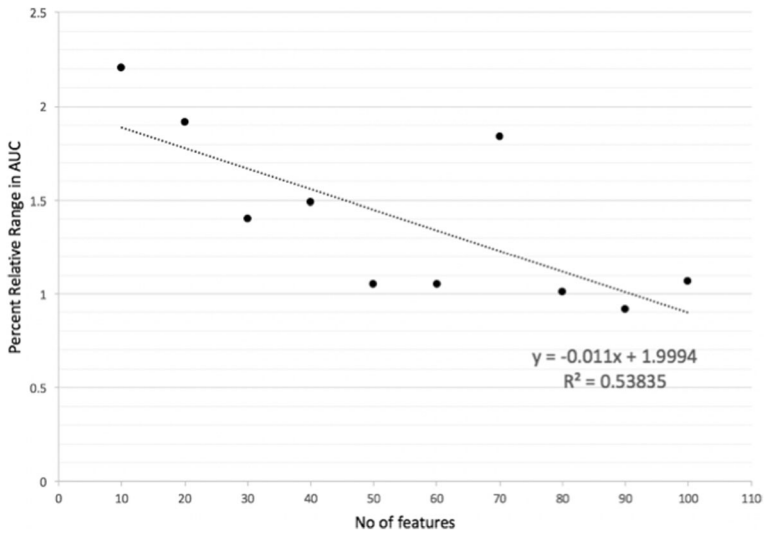

Fig. 3 Relationship between the number of selected MV features and sensitivity of MV models to changes in selection probability distribution P in terms of percent relative range in AUC

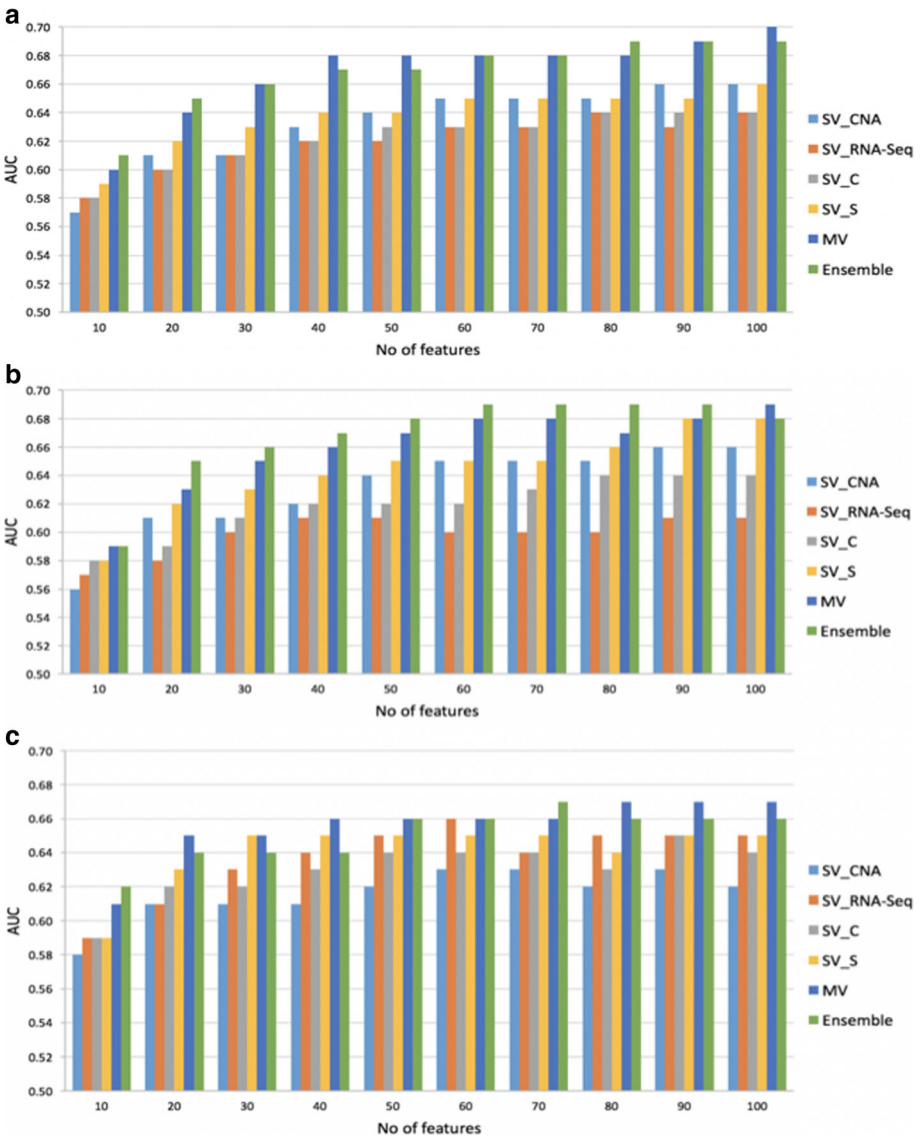

Fig. 4 Performance comparisons of final multi-view models with their single-view counterparts, for three different choices of machine learning algorithms: a) RF, b) XGB, and c) LR

Analysis of the top selected multi-view features

In order to get insights into the most discriminative features selected by our framework, we considered the top 100 features selected using MRMR-mv jointly from CNA and RNA-Seq views. To determine which features (genes) could serve as potential biomarkers for ovarian cancer survival, at each of the 50 iterations (resulting from running 5-fold procedure for 10 times), we scored each per-view input feature (input to our framework) by how many time it appears in the top 100 features. Table 6 summarizes the top 20 features from each view along with their normalized feature importance scores.

To examine the interplay between the top selected features from each view, we constructed an integrated network of interactions among the features using the cBio portal by integrating the biological interactions from public databases including NCI-Nature Pathway Interaction Database, Reactome, HPRD, Pathway Commons, and MSKCC Cancer Call Map [41]. Examination of the resulting network (Fig. 5) shows that *RPS19, PNOC*, *SFRP1* and *KCNJ16* are connected to other frequently altered genes, including *MYC or EIF3E* as oncogenes, from TCGA ovarian cancer dataset. In particular, ribosomal protein S19 (*RPS19*), which is known to be up-regulated in human ovarian and breast cancer cells and released from apoptotic tumor cells, was found to be associated with a novel immunosuppressive property [42]. Furthermore, *HTR3A* is targeted by several FDA approved cancer

Table 6 Top 20 features selected from CNA and RNA-Seq views

CNA	Score	RNA-Seq	Score
TBX18	0.44	OVGP1	0.56
TSHZ2	0.42	TOX3	0.54
RN7SL781P	0.42	SIX3	0.52
MAN1A2	0.42	HTR3A	0.50
KIF13B	0.40	FLG	0.48
DKFZP667F0711	0.36	SOSTDC1	0.48
CD70	0.36	EPYC	0.48
PRDM1	0.36	OBP2B	0.48
ZNF471	0.34	FBN3	0.46
RPS19	0.34	COL6A6	0.46
snoU13	0.34	NKAIN4	0.46
IRX1	0.32	LY6K	0.44
MIA	0.32	FABP6	0.44
LYPLA1	0.30	KIF1A	0.44
SHROOM3	0.30	KCNJ16	0.44
USP13	0.30	PNOC	0.42
SFRP1	0.28	TKTL1	0.42
CYP11A1	0.28	HLA-DRB6	0.42
ZMYM4	0.28	KRT14	0.42
APCDD1L	0.28	DPP10	0.40

drugs retrieved from PiHelper [43], an open source compilation of drug-target and antibody-target associations derived from several public data sources.

Finally, we performed a gene-set enrichment analysis to identify overrepresented GO terms in the two sets of top 20 features from CNA and RNA-Seq views. Specifically, we used the gene-batch tool in GOEAST (Gene Ontology Enrichment Analysis Software Toolkit) [44] with default parameters to import the gene symbols and to identify significantly overrepresented GO terms, for Biological Processes, Cellular Components and Molecular Function categories, in the CNA and RNA-Seq features sets. We found that the selected CNA gene set was enriched with 220 GO terms whereas the selected RNA-Seq gene set was enriched with 40 GO terms (See Additional files 1 and 2). Analysis of the GO terms enriched in the CNA gene set showed a significant overrepresentation of the molecular function GO terms related to hydrolase activity, oxidoreductase activity, and ion binding. Analysis of the GO terms enriched in the RNA-Seq gene set showed a significant over-representation of the molecular function GO terms related to transmembrane and substrate-specific transporter activity. We also used the Multi-GOEAST tool to compare the results of enrichment analysis of CNA and RNA-Seq gene sets. The graphical outputs of the Multi-GOEAST analysis results for top selected genes in CNA and RNA-Seq in Biological Processes, Cellular Components and Molecular Function categories are provided in Additional files 3, 4 and 5. In these graphs, red and green boxes represent enriched GO terms only found in CNA and RNA-Seq, respectively. Yellow boxes represent commonly enriched GO terms in both sets of genes. The saturation degrees of all colors represent the significance of enrichment for corresponding GO terms. Interestingly, GO:0003777~microtubule motor activity term is only shared GO term between CNA and RNA-Seq enriched terms (see Additional file 5). We concluded that the CNA and RNA-Seq features selected by the proposed multi-view feature selection algorithm are non-redundant not only in terms of the genes selected from the CNA and RNA-Seq views but also in terms of their significantly overrepresented GO terms.

Discussion

We presented a two-stage feature selection framework for multi-omics data integration. The proposed framework can be customized in different ways to implement a variety of data integration methods. We described a novel instantiation of the proposed framework using multi-view feature selection. We introduced MRMR-mv, which extends MRMR, one of the state-of-the-art single-view feature selection methods, to the multi-view setting. We used the proposed two-stage framework to conduct a set of experiments to compare the performance of single-view and multi-view methods for predicting ovarian cancer survival from multi-omics data. The results of our experiments demonstrate the potential of the two-stage feature selection framework in general, and the MRMR-mv multi-view feature selection method in particular, in integrative analyses of and predictive modeling from multi-omics data.

Evaluation of single-view models for predicting ovarian cancer survival using methylation data alone showed very poor predictive performance where as those trained using CNA or RNA-Seq data showed substantially better predictive performance (with AUC between 0.64 and 0.66). Multi-view models that integrate mult-omics data using MRMR-mv, a multi-view feature selection method, were able to outperform single-view models. For example, multi-view models using the top 100 features selected by MRMR-mv from CNA and RNA-Seq data were able to achieve an AUC of 0.7. With the anticipated rapid increase in the size of multi-omics data, we can expect the predictive performance of such models to show corresponding improvements.

Further improvements can be expected from better techniques for coping with the ultra high-dimensionality and sparsity of multi-omics data. Of particular interest in this context are methods for pan-cancer analysis [45], multi-task learning [46], and incomplete multi-view learning [47], and multi-view representation learning [23].

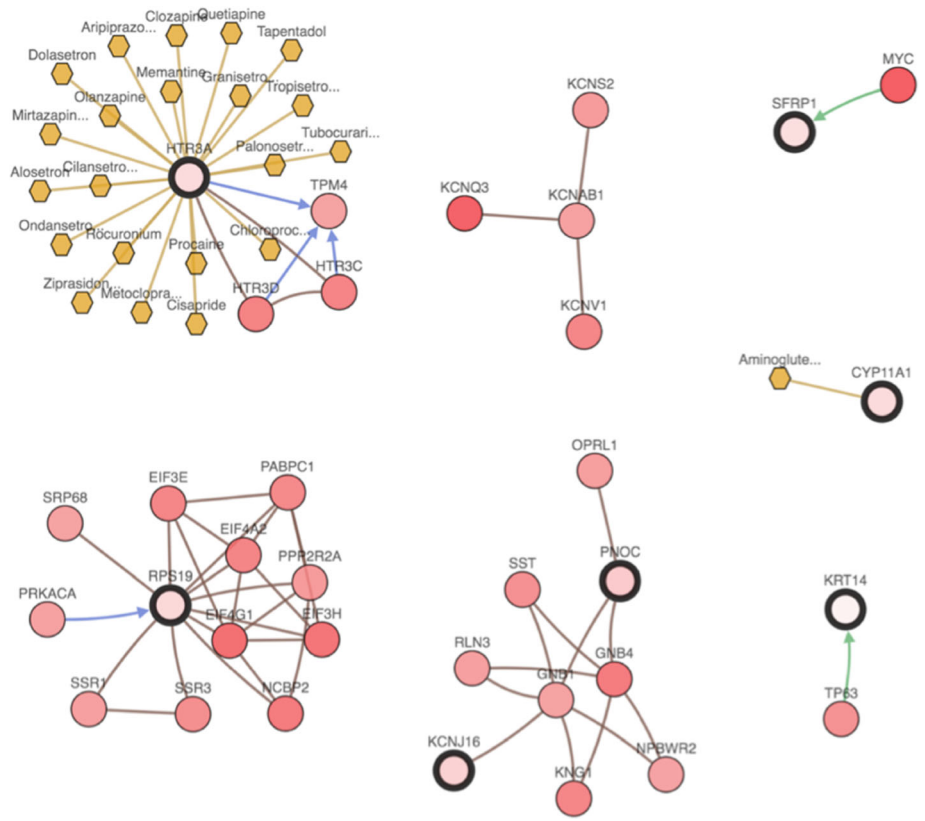

Fig. 5 Integrative network view of the features selected from CNA and RNA-Seq views. The genes corresponding to the selected features are highlighted by a thicker black outline. The rest of the nodes correspond to the genes that are frequently altered and are known to interact with the highlighted genes (based on publicly available interaction data. The nodes are gradient color-coded according to the alteration frequency based on CNA and RNA-Seq data derived from the TCGA ovarian cancer dataset via cBio Portal

MRMR-mv jointly selects (from multiple views) a compact yet most relevant subset of non-redundant features across multiple views for the prediction task at hand. Interestingly, the gene-set enrichment analysis of the top 20 genes selected by MRMR-mv from the CNA and RNA-Seq data shows that these genes are also non-redundant with respect to the GO terms that are significantly overrepresented in the CNA and RNA-Seq gene sets. If this observation is validated using other multi-omics datasets, MRMR-mv could be used to uncover, from multi-omics data, the underlying functional sub-networks that collectively orchestrate the biological processes that drive the onset and progression of diseases such as cancer. Ultimately, accurate and personalized prediction of clinical outcomes of different interventions and promising therapeutic targets for different cancer types will require advances in multi-view and multi-scale modeling that bring together information from different complementary data sources into cohesive explanatory, predictive, and causal models [48].

Conclusions

Developing multi-omics data-driven machine learning models for predicting clinical outcome, including cancer survival, is a promising cost-effective computational approach. However, the heterogeneity and extreme high-dimensionality of omics data present significant methodological challenges in applying the state-of-the art machine learning algorithms to training such models from multi-omics data. In this paper, we have described, to the best of our knowledge, the first attempt at applying multi-view feature selection to address these challenges. We have introduced a two-stage feature selection framework that can be easily customized to instantiate a variety of approaches to integrative analyses and predictive modeling from multi-omics data. We have proposed MRMR-mv, a novel maximum relevance and minimum redundancy based multi-view feature selection algorithm. We have applied the resulting framework and algorithm to build predictive models for ovarian cancer survival using multi-omics data derived from the Cancer Genome Atlas (TCGA).

We have demonstrated the potential of integrative analysis and predictive modeling of multi-view data in ovarian cancer survival prediction. Work in progress is aimed at further developing effective computational and statistical methods and tools for integrative analyses and modeling of multi-omics data, with particular emphasis on precision health applications.

Additional files

Additional file 1: GOEAST gene-batch output of enriched GO terms in the Biological Processes, Cellular Components and Molecular Function categories for CNA top selected genes. (TXT 45 kb)

Additional file 2: GOEAST gene-batch output of enriched GO terms in the Biological Processes, Cellular Components and Molecular Function categories for RNA-Seq top selected genes. (TXT 8 kb)

Additional file 3: Graphical output of Multi-GOEAST analysis results of Biological Processes GO terms in the top selected genes in CNA and RNA-Seq. (PDF 110 kb)

Additional file 4: Graphical output of Multi-GOEAST analysis results of Cellular Components GO terms in the top selected genes in CNA and RNA-Seq. (PDF 56 kb)

Additional file 5: Graphical output of Multi-GOEAST analysis results of Molecular Function GO terms in the top selected genes in CNA and RNA-Seq. (PDF 58 kb)

Abbreviations

AUC: Area under ROC curve; CNA: Copy number alteration; CV: Cross-validation; F-Stat: F-Statistic; GE: Gene expression; GOEAST: Gene Ontology Enrichment Analysis Software Toolkit; ICGC: International Cancer Genome Consortium; LR: Logistic Regression; MI: Mutual Information; MRMR: Min-Redundancy and Maximum-Relevance; MV: Multi-view; RF: Random Forest; RFE: Recursive Feature Elimination; TCGA: The Cancer Genome Atlas; XGB: eXtreme Gradient Boosting

Acknowledgments

We gratefully acknowledge the TCGA Consortium and all its members for the TCGA Project initiative, for providing sample, tissues, data processing and making data and results available. The results published here are in whole or part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions that constitute the TCGA research network can be found at http://cancergenome.nih.gov.

Funding

This project was supported in part by grants the National Institutes of Health (through grants NCATS UL1 TR000127, NCATS TR002014, NIGMS P50GM115318, NLM R01 NL012535 and NLM T32LM012415), the Pennsylvania Department of Health (SAP 4100070267), and the National Science Foundation (IIS 1636795). The project was also supported by the Edward Frymoyer Endowed Professorship in Information Sciences and Technology at Pennsylvania State University and the Sudha Murty Distinguished Visiting Chair in Neurocomputing and Data Science at the Indian Institute of Science [both held by Vasant Honavar] and the Pennsylvania State University Center for Big Data Analytics and Discovery Informatics (CBDADI) which is co-sponsored by the Institute for Cyberscience, the Huck Institutes of the Life Sciences, and the Social Science Research Institute at the university. The content, including specifically, analyses, interpretation, and conclusions, is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, NSF, or the Pennsylvania Department of Health or other sponsors. The publication costs were covered by CBDADI.

Availability of data and materials

The processed TCGA datasets used for analysis are publicly available at https://xenabrowser.net/datapages/

About this supplement

This article has been published as part of *BMC Medical Genomics* Volume 11 Supplement 3, 2018: Selected articles from the 7th Translational Bioinformatics Conference (TBC 2017): medical genomics. The full contents of the supplement are available online at https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-11-supplement-3.

Authors' contributions

YE, DK, and VH designed and conceived the research. YE developed and implemented the feature selection algorithm and the data integration framework. YE and TH ran the experiments. YE, TH, and MS performed the data analysis. YE drafted the manuscript. DK and VH edited the manuscript. All authors read and approved the final version of the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no conflict of interest.

Publisher's Note

Author details

[1]Artificial Intelligence Research Laboratory, College of Information Sciences and Technology, Pennsylvania State University, University Park, PA 16802, USA. [2]Biomedical and Translational Informatics Institute, Geisinger Health System, Danville, PA, USA. [3]The Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802, USA. [4]School of Electrical Engineering and Computer Science, Pennsylvania State University, University Park, PA 16802, USA. [5]The Center for Big Data Analytics and Discovery Informatics, Pennsylvania State University, University Park, PA 16802, USA. [6]The Clinical and Translational Sciences Institute, Pennsylvania State University, University Park, PA 16802, USA.

Published: 14 September 2018

References

1. Gagan J, Van Allen EM. Next-generation sequencing to guide cancer therapy. Genome Med. 2015;7(1):80.
2. Cancer Genome Atlas Research N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455(7216):1061–8.
3. Hudson TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, Bhan MK, et al. International network of cancer genome projects. Nature. 2010;464(7291):993–8.
4. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci. 2013;110(11):4245–50.
5. 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.
6. Gligorijević V, Malod-Dognin N, Pržulj N. Integrative methods for analyzing big data in precision medicine. Proteomics. 2016;16(5):741–58.
7. Network CGAR. Integrated genomic and molecular characterization of cervical cancer. Nature. 2017;543(7645):378.
8. Kim D, Shin H, Sohn KA, Verma A, Ritchie MD, Kim JH. Incorporating interrelationships between different levels of genomic data into cancer clinical outcome prediction. Methods. 2014;67(3):344–53.
9. Hanash S. Integrated global profiling of cancer. Nat Rev Cancer. 2004;4(8): 638–44.
10. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet. 2015;16(2):85–97.
11. Lussier YA, Li H. Breakthroughs in genomics data integration for predicting clinical outcome. J Biomed Inform. 2012;45(6):1199–201.

12. Kim D, Joung JG, Sohn KA, Shin H, Park YR, Ritchie MD, Kim JH. Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction. J Am Med Inform Assoc. 2015;22(1):109–20.
13. Kim D, Li R, Lucas A, Verma SS, Dudek SM, Ritchie MD. Using knowledge-driven genomic interactions for multi-omics data analysis: metadimensional models for predicting clinical outcomes in ovarian carcinoma. Journal of the American Medical Informatics Association. 2016; ocw165
14. Serra A, Fratello M, Fortino V, Raiconi G, Tagliaferri R, Greco D. MVDA: a multi-view genomic data integration methodology. BMC bioinformatics. 2015;16(1):261.
15. Kristensen VN, Lingjærde OC, Russnes HG, Vollan HKM, Frigessi A, Børresen-Dale A-L. Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer. 2014;14(5):299.
16. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. Front Genet. 2017;8:84.
17. Zhao J, Xie X, Xu X, Sun S. Multi-view learning overview: recent progress and new challenges. Information Fusion. 2017;38:43–54.
18. Huang C, Chung FL, Wang S. Multi-view L2-SVM and its multi-view core vector machine. Neural Netw. 2016;75:110–25.
19. Peng J, Aved AJ, Seetharaman G, Palaniappan K. Multiview boosting with information propagation for classification. IEEE Transactions on Neural Networks and Learning Systems. 2017;
20. Cai X, Nie F, Huang H: Multi-view k-means clustering on big data. In: Twenty-Third International Joint conference on artificial intelligence: 2013; 2013.
21. Chaudhuri K, Kakade SM, Livescu K, Sridharan K. Multi-view clustering via canonical correlation analysis. In: Proceedings of the 26th annual international conference on machine learning: 2009: ACM; 2009. p. 129–36.
22. Yang W, Gao Y, Shi Y, Cao L. MRM-lasso: a sparse multiview feature selection method via low-rank analysis. IEEE transactions on neural networks and learning systems. 2015;26(11):2801–15.
23. Wang W, Arora R, Livescu K, Bilmes J. On deep multi-view representation learning. In: International Conference on Machine Learning. 2015;2015:1083–92.
24. Goldman M, Craft B, Swatloski T, Cline M, Morozova O, Diekhans M, Haussler D, Zhu J. The UCSC cancer genomics browser: update 2015. Nucleic acids research. 2014; gku1073
25. Liu H, Motoda H: Feature selection for knowledge discovery and data mining, vol. 454: Springer Science & Business Media; 2012.
26. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinforma Comput Biol. 2005;3(02):185–205.
27. El Akadi A, Amine A, El Ouardighi A, Aboutajdine D. A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. Knowl Inf Syst. 2011;26(3):487–500.
28. Sakar O, Kursun O, Seker H, Gurgen F. Prediction of protein sub-nuclear location by clustering mRMR ensemble feature selection. In: Pattern Recognition (ICPR), 2010 20th International Conference on: 2010: IEEE; 2010. p. 2572–5.
29. Direito B, Duarte J, Teixeira C, Schelter B, Le Van Quyen M, Schulze-Bonhage A, Sales F, Dourado A. Feature selection in high dimensional EEG features spaces for epileptic seizure prediction. IFAC Proceedings Volumes. 2011; 44(1):6206–11.
30. Liu L, Cai Y, Lu W, Feng K, Peng C, Niu B. Prediction of protein–protein interactions based on PseAA composition and hybrid feature selection. Biochem Biophys Res Commun. 2009;380(2):318–22.
31. Zhang L, Zhang Q, Zhang L, Tao D, Huang X, Du B. Ensemble manifold regularized sparse low-rank approximation for multiview feature embedding. Pattern Recogn. 2015;48(10):3102–12.
32. Svd W, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. Computing in Science & Engineering. 2011;13(2):22–30.
33. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol. 1996:267–88.
34. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12(Oct):2825–30.
35. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
36. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2016. p. 785–94.
37. Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. Appl Stat. 1992:191–201.
38. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005;67(2):301–20.
39. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46(1):389–422.
40. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 1997;30(7):1145–59.
41. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401–4.
42. Markiewski MM, Vadrevu SK, Sharma SK, Chintala NK, Ghouse S, Cho J-H, Fairlie DP, Paterson Y, Astrinidis A, Karbowniczek M. The ribosomal protein S19 suppresses antitumor immune responses via the complement C5a receptor 1. J Immunol. 2017;198(7):2989–99.
43. Aksoy BA, Gao J, Dresdner G, Wang W, Root A, Jing X, Cerami E, Sander C. PiHelper: an open source framework for drug-target and antibody-target data. Bioinformatics. 2013;29(16):2071–2.
44. Zheng Q, Wang X-J. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. Nucleic acids research. 2008;36(suppl_2): W358–63.
45. Lengerich B, Aragam B, Xing EP. Personalized Regression Enables Sample-Specific Pan-Cancer Analysis. bioRxiv. 2018; 294496
46. Li Y, Wang J, Ye J, Reddy CK. A multi-task learning formulation for survival analysis. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 2016: ACM; 2016. p. 1715–24.
47. Xu C, Tao D, Xu C. Multi-view learning with incomplete views. IEEE Trans Image Process. 2015;24(12):5812–25.
48. Honavar VG, Hill MD, Yelick K: Accelerating science: a computing research agenda. arXiv preprint arXiv:160402006 2016.