Multi-omics Data and Analytics Integration in Ovarian Cancer

Archana Bhardwaj, Kristel Van Steen

Multi-omics data and analytics integration in ovarian cancer

Archana Bhardwaj[1], Kristel Van Steen[1]

[1]GIGA-R Centre, BIO3 – Medical Genomics,
University of Liège, Liège, Belgium
a.bhardwaj@uliege.be

Abstract. Cancer, which involves the dysregulation of genes via multiple mechanisms, is unlikely to be fully explained by a single data type. By combining different "omes", researchers can increase the discovery of novel bio-molecular associations with disease-related phenotypes. Investigation of functional relations among genes associated with the same disease condition may further help to develop more accurate disease-relevant prediction models. In this work, we present an integrative framework called Data & Analytic Integrator (DAI), to explore the relationship between different omics via different mathematical formulations and algorithms. In particular, we investigate the combinatorial use of molecular knowledge identified from omics integration methods netDx, iDRW and SSL, by fusing the derived aggregated similarity matrices and by exploiting these in a semi-supervised learner. The analysis workflows were applied to real-life data for ovarian cancer and underlined the benefits of joint data and analytic integration.

Keywords: Multi-omics integration, Semi-supervised learning, Network medicine.

1 Introduction

Worldwide, ovarian cancer has the worst prognosis and highest mortality rate [1]. Coupling biomarker discovery to survival traits can increase our understanding about relevant tumor mechanisms and may provide insights into early detection strategies and/or preventive actions. The abundance of data due to advancements in high throughput sequencing technologies and carefully established data repositories are essential in this context. Cancer biology is complex and requires systems views to unravel the complexity. One of the Big Data cancer repositories are made available via The Cancer Genome Atlas Program (TCGA - https://www.cancer.gov/about-nci/organization/ccg/re-

search/structural-genomics/tcga). It comprises multiple omics collections such as transcriptome, methylation and copy number variant (CNV) data. A transcriptome refers to the full range of messenger RNA that is produced in a particular cell or tissue type. A methylome, giving rise to methylation data, comprises the set of all nucleic acid methylation modifications in the genome of an organism or in a particular cell. CNVs are a specific type of DNA variation referring to copies of sections of the genome, the number of which varying between individuals. Even though non-omics data should not be ignored, in general, adopting multi-omics integrative strategies in cancers, like ovarian cancer, are believed to be the road to travel by, irrespective of whether subtyping or (survival) prediction is the aim (e.g., [2-3]). With the vast amount of data to be mined, it is not surprising that machine learning tools have become indispensable in the data integration field, including multi-view methods for joint clustering of multiple data types [4], auto-encoder architectures based on omics and clinical data to study a variety of cancer-relevant traits [5], and deep-learners for robust cancer survival prediction [3].

While performing multi-omics integration, several challenges exist, such as validating the added value of multi-omics data integrative methods over single-omics analyses, assessing at which stage to perform the integration (e.g., early – data integration before analytic modelling, late – integration of modelling results), and how to deal with concordant and discordant relationships between multi-omics datasets in cancers. Here, we explore the performance of a novel combined omics data and analytics integrator (DAI) and compare it to state-of-the-art multi-omics data integrative approaches. We define performance in terms of optimized prediction or classification of ovarian cancer patients into short-term (less than 3 years) or long-term survival (at least 3 years). The categorization based on the threshold of 3 years of survival was inspired by [6]. We consider 3 omics data types: genomic (CNVs), epigenomic (methylation) and transcriptomic (gene expression). Notably, epigenomics refers to "epi"-genetic ("epi" from Greek: on top of) modifications that affect gene expression regulation but does not change the genomic sequence itself.

The paper is organized as follows. In Section 2, data overview and preparation steps are outlined. Analytical workflows are detailed in Section 3. Results are presented in Section 4. A discussion and closing remarks are given in Section 5.

2 Data Overview and Preparation

CNV, methylation and gene expression data for ovarian cancer were retrieved from the TCGA data portal via TCGA2STAT [7] software. In particular, we first discarded patients who did not have the 3 omics data types available. We then used the OMICSBind function TCGA2STAT to merge the available data and subsequently performed sample filtration following [8]. OMICSBind returns a combined data matrix for samples that are common to two types of molecular input data. Thereafter, we discarded patients having "vital status" as "dead" and "days to death" as "non-positive" or "NA", and we

discarded patients having “vital status” as “alive” and “days to last follow-up” as “non-positive” or “NA” [8]. Next, we created two groups: ST (< 3 years of survival) and LT (≥ 3 years of survival). Based on the above filtration criteria, LT/ST status was available for all patients included in this study (i.e. no missing labels). For each data type, we eliminated genes with a missing rate across all samples >20%. Remaining missing omics data entries were imputed with the kNNImpute function in R. In particular, each missing feature for a sample was replaced by a weighted average of the corresponding features from k nearest neighbors of that sample, weighted by the distance of the neighbors [9]. The resulting dataset for integrative analyses comprised 100 ST and 130 LT survivors, with information available on 22618 CNVs, 12644 methylation and 12043 gene expression features.

3 Analytical workflows

As the aim is to optimize classification/prediction of LT/ST survival status and to exploit the integrated information of 3 omics datasets, we used the following promising integrative approaches as starting point: iDRW [10], netDx [11], and SSL [6], with default options, unless specified otherwise. Each of these methods adopt different paths towards generating omics features, that is the basis to assess similarities between patients. Apart from applying the original work-flows, patient-similarity matrices obtained from each approach were fused (when applicable) to create a single matrix per method, which was submitted to a graph-based learning method as in [6], so as to classify patients into LT/ST survival groups (Fig. 1). More details are given in the following paragraphs.

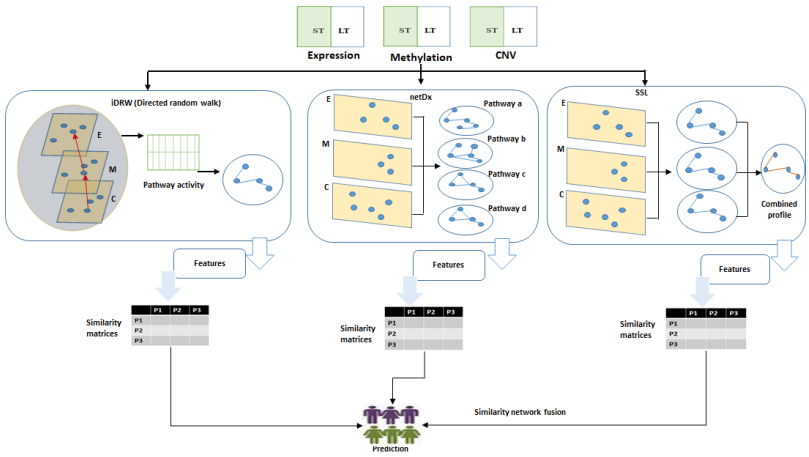

Fig. 1. DAI workflow to create data and analytics integrated patient similarity networks using adaptions of the machine learning approaches iDRW, netDx and SSL (see text). We employed

molecular information on expression (E), methylation (M) and CNV (C) data and created new features on the basis of which to assess patient similarity. For iDRW an integrated omics network was used to derive pathway activity scores as new features. In netDx, features were pathway-genes and omics-specific patient similarity networks were derived for each pathway (4 are shown). Linked to SSL, original gene measurements (data-driven) and specific knowledge-based gene sets (knowledge-driven) features that carry information about disease relevance and protein networks were used. Developed patient similarity networks were combined into a method-specific single network. Principles of similarity network fusion were used to generate an analytics integrated patient similarity network, which served as input to a semi-supervised learning method to predict LT/ST survival state.

3.1 Integrative Directed Random Walk-based workflow (iDRW) [10]:

In this approach gene-gene networks are built for each omics dataset, supported by KEGG pathway information (https://www.genome.jp/kegg/), from which an integrated directed network is derived. Then a random walk is performed on this integrated network. Significant genes in the integrated gene-gene networks and their weights from the DRW method contribute to integrated pathway-activity scores [10]. For our purposes we used a customized version from the authors to handle >2 omics. Out of 327, only the six significant pathways (T-test of pathway-activity across LT/ST survivor classes) were kept. Next, in line with the original iDRW workflow, a regression model in R was applied that classified the samples into ST and LT classes. As in the current study the focus lies on integration and not on variations of prediction model paradigms, we replaced the logistic regression model by a graph-based semi-supervised learner that can be applied with missing classes and with multiple input data types (i.e., iDRW+SSL). We thus converted the patient similarity matrix W_{iDRW} to a Laplacian L_{iDRW} and obtained final class predictions by solving $(I + \mu L_{iDRW})^{-1} y$, with y encoded as (-1,1) corresponding to (ST survival, LT survival) and μ a trade-off parameter, following the single-graph based semi-supervised learner of [6]. Note that in the presence of a missing survival status for a patient, it would suffice to encode the corresponding response y as 0.

3.2 Supervised patient classification algorithm via patient similarity networks (netDx) [11]:

The approach constructs patient-patient similarity networks for each gene set of interest per data types. As before, we used CNV, methylation and gene expression data. This is followed by a network selection (i.e feature selection) step based on the netDx scoring procedure. Here, netDx score for each feature (i.e pathway) indicates the number of times that feature was assigned a positive score in a query during resampling process. Scoring process was repeated for each class (ST and LT). At end, the best network is one for which edges only exist between individuals of the same class (e.g. LT survivor)

and not the other (e.g. ST survivor). An overall patient similarity network is subsequently created by integrating feature-selected networks (patient similarity matrix W_{netDx}). The original netDx strategy to predict survival status was compared to an adaption (netDX+SSL) using the semi-supervised learner as before with predicted classes obtained by solving $(I + \mu L_{netDx})^{-1} y$ (L_{netDx}: the graph Laplacian linked to W_{netDx}).

3.3 Graph-based semi-supervised learning (SSL) [6]:

Also here, the approach is based on creating patient similarity matrices for each omics data type separately. However, the features used to assess patient-to-patient similarity is different from the previous approaches. In particular, pre-defined gene sets as "genomic knowledge" were downloaded from the Molecular Signatures Database (MSigDB 7.0).32 [12]: chemical and genetic perturbations and canonical pathways (C2), motif (C3) and cancer gene sets (C4), gene ontology (C5), and immunological signatures (C7), involving 5501, 831, 858, 9996 and 4872 gene sets, respectively. We also collected a list of 2067 "seed genes" from the OCGene database, appended with genes from Papp et al. [13], leading to a unique seed gene list of 2072 genes. These were submitted to ToppGenet [14] to prioritize neighboring genes of the seeds based on functional similarity to the seeds or topological features in a protein-protein interaction network. The top 1% prioritized genes (1600 genes) were used to refine the MSigDB-derived "genomic knowledge" gene sets (number of genes in C2: 3132, C3: 568, C4: 449, C5: 5593, C7: 2711). Gene measurements per patient were subsequently averaged within each genomic knowledge gene set and were used to create "knowledge-driven" patient similarity matrices W_{c2_exp}, W_{c2_cnv}, W_{c2_meth} , W_{c3_exp}, W_{c3_cnv}, W_{c3_meth}, W_{c4_exp}, W_{c4_cnv}, W_{c4_meth}, W_{c5_exp}, W_{c5_cnv}, W_{c5_meth}, W_{c7_exp}, W_{c7_cnv}, and W_{c2_meth}. This is in contrast to using all original gene measurements, which would lead to "data-driven" similarity matrices W_{DD_E}, W_{DD_CNV}, W_{DD_METH}, for unfiltered measurements of gene expression, CNV and methylation, respectively. For each gene set of interest, the weights α for these matrices were estimated so as to optimize LT/ST survival class prediction as in the minimization problem $\min_{\alpha} y^T (I + \sum_{k=1}^{K} \alpha_k L_k)^{-1} y, \sum_k^K \alpha_k \leq \mu$ (K: number of graphs = 3; L_k: Laplacian corresponding to graph k; y: class response vector). The final class predictions were obtained by $(I + \sum_{k=1}^{K=3} \alpha_k L_k)^{-1} y$.

3.4 Data & Analytics Integrator (DAI):

We started by adapting netDx as follows. We obtained a single similarity matrix by fusing ST and LT specific similarity matrices. In particular, multiple pathway profiles

for ST patients were integrated while adopting Similarity Network Fusion (SNF) analytics [15], leading to $W_{netDx-ST}$, and similar for LT patients, leading to an aggregated patient similarity matrix $W_{netDx-LT}$. The fused matrix was denoted by W_{netDx_S}, where the underscore "S" now refers to the pooled LT and ST survivors. This matrix was subsequently converted to a Laplacian L_{netDx_S} for use in the semi-supervised learner of [6], as explained before.

Then SSL was adapted to generate a single patient similarity matrix by first retrieving the software's weights α_i for each data type i (expression, methylation, and CNV). Second, we normalized the retained weights (i.e. new weights sum to 1) to form an integrated patient similarity matrix $W_{SSL\_adapted} = \sum_i \alpha_i M_i$, with Mi denoting the patient similarity matrix derived from omics data type i. Lastly, we built a shell around adaptations of iDRW, netDx and SSL integrating all three matrices W_{iDRW}, W_{netDx_S} and $W_{SSL\_adapted}$. In particular, we SNF-fused the three matrices and converted the resulting data and analytics aggregated patient similarity matrix W_{DAI} to a Laplacian L_{DAI}. Class predictions were obtained by solving $(I + \mu L_{DAI})^{-1} y$, with y encoded as (-1,1,0) corresponding to (ST survivor, LT survivor, unknown survival state), and with μ a trade-off parameter between predictions close to the given label and predictions not too different from those for graph-adjacent nodes. DAI (Fig. 1.) also allows the option to apply the multi-graph based SSL model of [6] on L_{iDRW}, L_{netDx} and L_{SSL} directly, instead of first fusing patient similarity matrices and second applying a single-graph semi-supervised learner. Key (dis-) similarities between DAI and iRDW, netDx and SSL in their original forms are summarized in Table 1.

Table 1. Highlighted (dis-)similarities between DAI and original implementations of iDRW, netDx and SSL.

Highlight	**SSL**	**netDx**	**iDRW**	**DAI**
Input data (Omics based)	Data-driven or knowledge-driven	Data driven	Data driven	Knowledge-driven
Weighted data accommodated	Yes	No	No	Yes
Feature selection	No additional feature selection	Identify omics-specific features leading to pathway-specific profiles	Identify genes contributing to pathway activity score	No additional feature selection
Output includes patient similarity matrix	Yes	Yes	No	Yes
Prediction model via multi/single graph-based semi-supervised learning	Yes	No	No	Yes

4 Results

4.1 Single Omics analyses

We first analyzed each omics data type separately, in a data-driven and knowledge-drive fashion, as explained in Section 3 (SSL). For the ovarian samples, W_{c2_cnv}, W_{c3_cnv}, W_{c4_cnv}, W_{c5_cnv} and W_{c7_cnv}, typically gave rise to the highest AUC values (Fig. 2). Overall, knowledge-based SSL outperformed data-driven SSL. Single graph-based SSL based on patient similarity for C2, C3, C4, C5, and C7 typically increased AUC estimates compared to data-driven approaches.

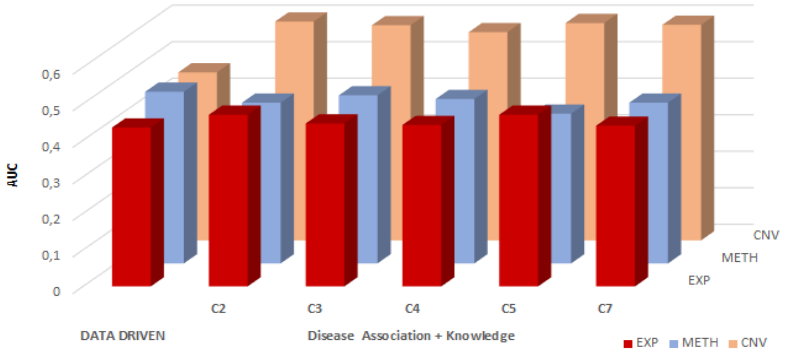

Fig. 2. Single-omics prediction performance. By data type AUC estimates for data-driven (W_{DD_exp}, W_{DD_cnv}, W_{DD_meth}) and knowledge-driven models (W_{c2_exp}, W_{c2_cnv}, W_{c2_meth} , W_{c3_exp}, W_{c3_cnv}, W_{c3_meth}, W_{c4_exp}, W_{c4_cnv}, W_{c4_meth}, W_{c5_exp}, W_{c5_cnv}, W_{c5_meth}, W_{c7_exp}, W_{c7_cnv}, and W_{c7_meth}).

4.2 Multi Omics integration

Next, based on 3-omics integration with iDRW, we identified numerous significant genes. Using the original workflow of iDRW, a total number of 1145, 2544, and 1846 genes from CNV, methylation, and expression omics, respectively, were found to be uniquely significant (unadjusted p value < 0.05). We mapped all these genes on their respective chromosomes (Fig. 3A: circular plot). Only 32 were common (*ABCA2, ACSL3, AKAP1, ALPI, AP3B2, APTX, ARPC2, CD79B, CLTC, COLEC12, COX11, CSHL1, CYB561, DDX42, EXOSC9, FAm83E, HOXB9, INTS9, LEPROTL1, mPHOSPH10, NFE2L1, NR2F2, OSBPL7, PmL, RPP25, SNF8, SNRPG, TIA1, TJP1, TUBD1, UBAC1, ZNF652*). Characteristic for these common genes was that they appeared to be highly co-expressed (Fig. 3B). iDRW's multi-omics view highlighted 6 statistically significant pathways, implying that their corresponding pathway scores

were significantly different between LT/ST survivors. The AUC of the original iDRW using a logistic regression model was estimated to be 0.32, which is lower than AUC=0.51 with our adapted version using W_{iDRW} and predictions based on single graph semi-supervised learning.

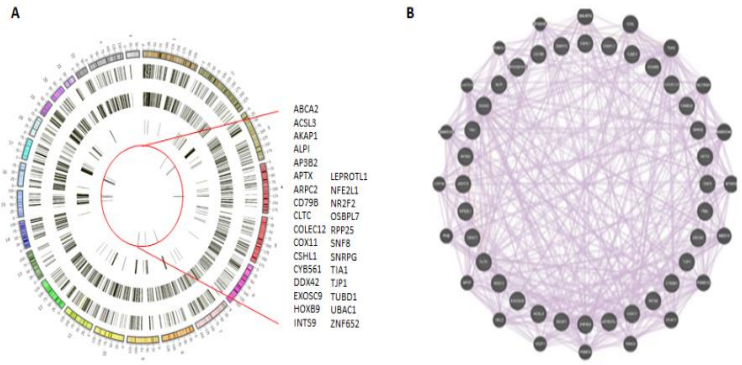

Fig. 3. iDRW identified significant genes. A. Circos plot: Chromosomal distribution of significant genes unique to either CNV, methylation, and expression omics. The outer circle represents the chromosomal bands. First inner circles represent the distribution of omics specific (inward direction: CNV, methylation, and expression) genes. The red circle (i.e. fourth layer) encapsulates significant genes common across three omics after integration; B. Gene-expression network of 32 significant genes, common to all considered ovarian omics data types.

Application of netDx to ovarian patient samples showed a higher number of KEGG pathways crossing the netDx threshold criterion for LT/ST survival prediction (Fig. 4). These pathways profiles were converted into patient specific similarity matrices to derive group specific W_{netDx_ST} and W_{netDx_LT} weight matrices (see also Section 3).

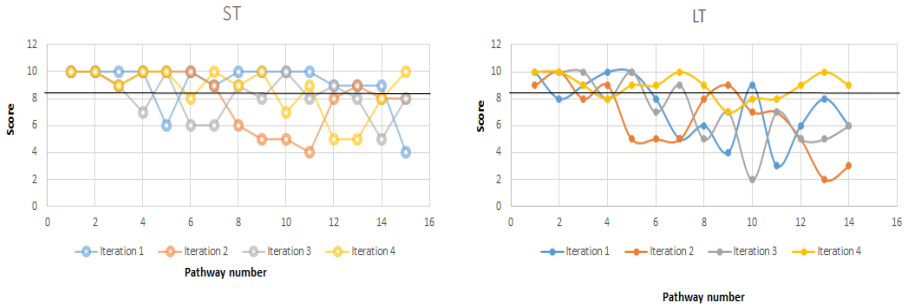

Fig. 4. Feature selection in netDx. Scatter plot to indicate the score of the 15 pathways across four different runs (iteration) in ST and LT survivor patients. The horizontal line is the netDx proposed threshold.

Furthermore, the original netDx implementation gave AUC=0.50 (Fig.5). The adapted version with SNF fused similarity matrix W_{netDx_S} submitted to single-graph SSL [6] improved the performance (AUC=0.60). Multi-omics profiles were integrated across biological knowledge to increase prediction with SSL over a single omics approach. By integration of W_{c2_exp}, W_{c2_cnv}, and W_{c2_meth} we achieved an AUC of 0.52. The predication accuracies were quite similar for other sources of biological knowledge. In particular, integration of W_{c3_exp}, W_{c3_cnv}, W_{c3_meth}, led to AUC=0.51; integration of W_{c4_exp}, W_{c4_cnv}, W_{c4_meth}, W_{c5_exp}, W_{c5_cnv}, W_{c5_meth}, W_{c7_exp}, W_{c7_cnv}, an W_{c7_meth} resulted in AUC prediction accuracies of 0.51, 0.55, and 0.53, respectively.

The current test version of DAI can be seen as a simple wrapper approach around multiple data integrative analytics to increase class prediction. Rather than submitting W_{iDRW}, $W_{netDx\_S,}$ $W_{SSL\_adapted}$ to a multiple graph-based semi-supervised learner, we primarily focused on obtaining a fused similarity network and single graph-based learning. With this setting, DAI's estimated AUC of 0.51 clearly outperformed the original iDRW, yet showed comparable performance to the original implementations of netDx and SSL. Among the original implementations of iDRW, netDx and SSL multi-omics prediction strategies, iDRW was the worst performer (AUC= 0.32). Interestingly, our adapted version of netDx (i.e adapting the prediction model itself) outperformed all other considered strategies, including iDRW+SSL and DAI that performed similarly to the original SSL multi-omics integrative method (Fig. 5). To investigate whether there was an added value of multiple over single graph-based learning, combined with netDX and learning over 19 graphs, a smaller AUC was obtained compared to netDx+SSL (not shown), but DAI's performance (involving 3 graphs) increased to give the highest AUC (0.62) among all considered approaches (Fig. 5 – DAI*).

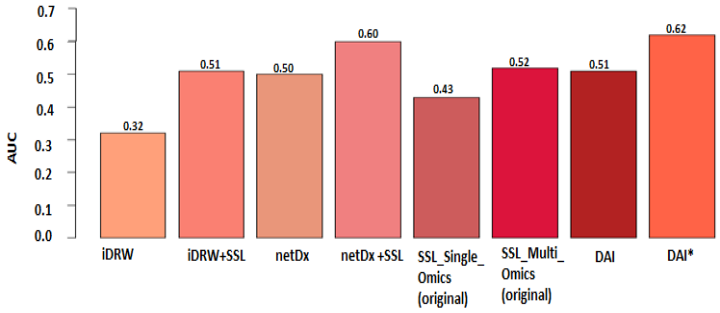

Fig. 5. Prediction performance of multiple data integrative analysis workflows and DAI. Prediction performance is measured by AUC. Legend: iDRW+SSL uses single-graph SSL to the integrated pathway-activity based patient similarity matrix; netDx+SSL applies multi-graph SSL to a similarity network fused matrix; SSL_Single_Omics (original) exploits a single-omics based knowledge-driven (pathways) patient similarity matrix and single-graph SSL, giving rise to an omics-specific AUC; AUCs are averaged across multi-omics;; SSL_Multi_Omics (original) employs knowledge-driven (pathways) patient similarity matrices across multiple omics combined with multi-graph SSL; DAI combines fusion of W_{iDRW}, W_{netDx_S} and $W_{SSL\_adapted}$ with single

graph-based SSL. DAI* differs from DAI in that W_{iDRW}, W_{netDx_S} and $W_{SSL\_adapted}$ are not fused but combined with multip-graph based SSL.

5 Discussion & final remarks

We introduced a Data and Analytic Integrator (DAI) that attempts to improve disease class prediction accuracy by integrating multi-omics data and analytics in various ways. The current workflow integrates 3 types of omics data, being CNV, methylation and gene expression data, and 3 analytic frameworks, represented by iDRW, netDx and SSL. Each of the analytics approaches derives information from multi-omics in a unique way and thus maximize their potential of providing complementary information towards class predication. In DAI, extracted information from each approach is translated into a single patient similarity matrix. The matrices for each of the analytic approaches are then combined. The current implementation of DAI uses Similarity Network Fusion to create an aggregated matrix but also allows using the individual matrices directly into a graph-based semi-supervised learner to predict class membership. The latter seems to be advantageous in terms of AUC performance, especially when the number of graphs for learning is relatively small. More work is needed though to investigate the impact of aggregating highly heterogeneous analytics.

As disease-associated genes are helpful in generating hypotheses about disease mechanisms, we investigated the utility of filtered gene sets, by making explicit use of earlier reported disease-gene associations. Little added value was achieved by doing so, compared to using unfiltered gene sets, except for giving rise to reduced computation times. One explanation may lie in the fact that association models and prediction models have different aims and evaluation criteria. Pathways highlighted by DAI (in particular RANBP2 pathways;SMARCA4 pathway; NOL7 pathways; diabetes pathways) were found to be implicated in ovarian, breast, cervical, and neuroblastoma cancer types [16-18].

In summary, our pilot results have shown that the exploitation of knowledge-based gene sets can substantially increase prediction performance. Furthermore, letting the data speak for themselves, in that the contribution of multiple omics data types in prediction models is estimated from the data, seems to boost prediction performance, but cannot receive all the credits. For instance, simply changing a logistic regression prediction model for a predictor based on a single aggregated patient similarity matrix was sufficient to create a top performer. Also, including a poor performer in similarity network fusion of three patient similarity matrices, based on multi-omics view from 3 analytic approaches, did not work decremental. Hence, future work will include the further exploitation of knowledge-driven data in DAI in combination with more elaborate non-linear aggregation of method-specific patient similarity matrices that estimate the relative contribution of each such matrix with the objective to maximize prediction accuracy.

Acknowledgement

We thank So Yeon Kim for sharing extended code of iDRW that allows the integration of 3 omics data types. A.B, and K.V.S acknowledge funding by Télévie 2015 “PDAC-xome: Exome sequencing in PDAC” (convention n° 7.4629.15), Télévie 2016 “Drivers and markers in pancreatic cancer” (convention n° 7.4502.16), and FRS-FNRS – CDR 2017 “SysMedPC” (convention n° J.0061.17).

References

1. Momenimovahed, Z., Tiznobaik, A., Taheri, S., Salehiniya, H.: Ovarian cancer in the world: epidemiology and risk factors. International journal of women’s health 11, 287-299 (2019).
2. Shen, R., Mo, Q., Schultz, N., Seshan, VE., Olshen, A.B., Huse, J. et al: Integrative subtype discovery in glioblastoma using iCluster. PLoS One 7(4), e35236 (2012).
3. Chaudhary K, Poirion O.B, Lu L, Garmire, L.X.: Deep learning–based multi-omics integration robustly predicts survival in liver cancer. Clinical Cancer Research 24(6), 1248–1259 (2008).
4. Rappoport, N., Shamir, R.: Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic acids research 46(20), 10546–10562 (2018).
5. Simidjievski, N., Bodnar ,C., Tariq, I., Scherer, P., Terre, HA., Shams, Z., et al.: Variational autoencoders for cancer data integration: design principles and computational practice. Frontiers in Genetics 10,1205 (2019).
6. Kim, D., Joung, J-G., Sohn, K-A., Shin, H., Park, YR., Ritchie, MD., et al.: Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction. Journal of the American Medical Informatics Association 22(1), 109–120 (2014).
7. Wan, Y.W., Allen, GI., Liu, Z.: TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. Bioinformatics 32(6), 952–954 (2016).
8. Dereli, O., Oğuz, C., Gönen, M.: Path2Surv: Pathway/gene set-based survival analysis using multiple kernel learning. Bioinformatics 35(24), 5137–5145 (2019).
9. Xu, A., Chen, J., Peng, H., Han, G., Cai, H.: Simultaneous interrogation of cancer omics to identify subtypes with significant clinical differences. Frontiers in genetics 10, 236 (2019).
10. Kim, S.Y., Jeong, H.H., Kim, J., Moon, J.H., Sohn K.A.: Robust pathway-based multi-omics data integration using directed random walks for survival prediction in multiple cancer studies. Biology direct 14(1), 8 (2019).
11. Pai, S., Hui, S., Isserlin, R., Shah, M.A., Kaka, H., Bader, G.D.: netDx: Interpretable patient classification using integrated patient similarity networks. Molecular systems biology 15(3), (2019).
12. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., Mesirov, J.P.: Molecular signatures database (MSigDB) 3.0. Bioinformatics 27(12), 1739–40 (2011).
13. Papp, E., Hallberg, D., Konecny, G.E., Bruhm, D.C., Adleff, V., Noë, M., et al.: Integrated genomic, epigenomic, and expression analyses of ovarian cancer cell lines. Cell reports 25(9), 2617–33 (2018).

14. Chen, J., Bardes, E.E., Aronow, B.J., Jegga, A.G.: ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic acids research 37(suppl_2), W305–311 (2009).
15. Wang, B., Mezlini, A.M, Demir, F., Fiume, M., Tu, Z., Brudno, M., et al.: Similarity network fusion for aggregating data types on a genomic scale. Nature methods 11(3), 333 (2014).
16. Connor, Y.D, Miao, D., Lin, D.I., Hayne, C., Howitt, B.E., Dalrymple, J.L, et al.: Germline mutations of SMARCA4 in small cell carcinoma of the ovary, hypercalcemic type and in SMARCA4-deficient undifferentiated uterine sarcoma: Clinical features of a single family and comparison of large cohorts. Gynecologic Oncology (2020).
17. Doçi, C.L, Mankame, T.P., Langerman, A., Ostler, K.R., Kanteti, R., Best, T., et al.: Characterization of NOL7 gene point mutations, promoter methylation, and protein expression in cervical cancer. International journal of gynecological pathology: official journal of the International Society of Gynecological Pathologists 31(1), 15-24 (2012).
18. Blanco Jr, L.Z., Kuhn, E., Morrison, J.C., Bahadirli-Talbott, A., Smith-Sehdev, A., Kurman, R.J.: Steroid hormone synthesis by the ovarian stroma surrounding epithelial ovarian tumors: a potential mechanism in ovarian tumorigenesis. Modern Pathology 30(4), 563–576 (2017).