Multi-*omics* integration for neuroblastoma clinical endpoint prediction

M. Francescatto[a], S. Rezvan Dezfooli[a], A. Zandonà[a,b,c], M. Chierici[a], G. Jurman[a], C. Furlanello[a,*]

[a] *Fondazione Bruno Kessler, Trento, Italy*
[b] *Centre for Integrative Biology, University of Trento, Italy*
[c] *Department of Information Engineering, University of Padova, Italy*

Abstract

Recent high-throughput methodologies such as microarrays and next-generation sequencing are well established / routinely used in cancer research, generating large amounts of complex data at different *omics* layers. The effective integration of *omics* data could provide a broader insight into the mechanisms of cancer biology, helping researchers and clinicians to develop personalized therapies. Here we explore the use of Integrative Network Fusion, a bioinformatics framework combining a novel similarity network fusion method and machine learning for the integration of multiple *omics* data. We apply the framework for the classification of neuroblastoma patients belonging to multiple disease stages integrating microarray, RNAseq and array comparative genomic hybridization data. We provide detailed results for one case study, the integration of microarray and CNV data for the classification and prediction of Event-Free Survival patients. Further, we discuss ongoing work on the dataset proposed for the CAMDA2017 neuroblastoma challenge.

1. Introduction

Neuroblastoma is a rare disease typically manifesting in early infancy with an estimated 700 new cases diagnosed in the US each year [1]. It displays a very heterogeneous clinical course, with extreme cases presenting spontaneous regression opposed by patients relapsing and eventually dying despite prompt therapy [2]. Because of this heterogeneity, the ability to accurately predict the most likely disease outcome at the time of diagnosis is of extreme importance, especially given that accurate risk estimation allows delivering an appropriate targeted therapy [3]. Amplification of the oncogene *MYCN* and age at diagnosis are currently the key clinical characteristics for the patient's risk assessment [4]. However these indicators only cover a portion of all neuroblastoma cases (ca. 22% of all neuroblastoma tumors present *MYCN* amplification [2]). The introduction of genome wide assays able to probe in great detail multiple genomics aspects often at affordable prices brought the promise of novel biomarkers identification for clinical outcome prediction, especially in combination with effective data analysis [5, 6]. Machine learning approaches have been adopted for the predictive classification of patient outcome in neuroblastoma, also considering the integration of data from multiple assays [5, 7]. For example, in a previous effort, the RNA Sequencing Quality Control (SEQC) initiative extensively explored expression-based predictive models for neuroblastoma risk assessment [8]. However, comprehensive integrative approaches effective across multiple clinical outcomes are still limited [5]. For this reason, we set out to apply Integrative Network Fusion (INF), a novel integrative approach able to combine multiple data types in a machine learning setting [9]. On the CAMDA2017 neuroblastoma challenge dataset (145 samples), INF improves prediction on Event-Free Survival (EFS) when combining microarray and comparative genomic hybridization array (aCGH) data with respect to both simple juxtaposition and use of independent datasets. Interestingly, the approach identifies a subset of samples that is consistently misclassified. For the remaining endpoints and on the full set of 498 samples, classification results are more heterogeneous, with classification performances displaying large variation across endpoints, as previously observed [8].

[*] Corresponding/presenting author

Email addresses: francescatto@fbk.eu (M. Francescatto), rezvan@fbk.eu (S. Rezvan Dezfooli), zandona@fbk.eu (A. Zandonà), chierici@fbk.eu (M. Chierici), jurman@fbk.eu (G. Jurman), furlan@fbk.eu (C. Furlanello)

Endpoint	498 cohort		145 cohort	
	TR	TS	TR	TS
ALL-EFS ALL-OS	249	249	71	74
CLASS	136	136	42	45
HR-EFS HR-OS	86	90	NA	NA

Table 1: Sample stratification (number of subjects). TR: training set; TS: test set; NA: not applicable (number of samples too low for accurate classification).

2. Materials and methods

Samples. The datasets used in this study include RNA-Seq and Agilent microarray gene expression profiles of 498 neuroblastoma patients [8], as well as matched aCGH data for a subset of 145 patients [10–13]. Clinical characteristics of the 498 samples were described previously [8]. The following prognostic endpoints were considered for classification tasks: the occurrence of an event (progression, relapse or death) (ALL-EFS); the occurrence of death from disease (ALL-OS); patient gender (SEX); an extreme disease outcome (CLASS); the occurrence of an event (HR-EFS) and death from disease (HR-OS) in high-risk (HR) patients only. HR status was defined according to the NB2004 risk stratification criteria. Samples were split in training (TR) and test (TS) sets according to previous partitioning [8]. Outcome stratification statistics are summarized in Table 1.

Data processing. Agilent microarrays (AG1) and RNA-Seq (MAV) were preprocessed elsewhere [8], obtaining log_2 normalized expressions for probes (AG1-P), genes (MAV-G), and transcripts (MAV-T). Here, AG1-P were further summarized over genes (AG1-G) according to platform annotation, while aCGH microarray raw data (CNV) were preprocessed by the rCGH R/Bioconductor package [14] using default parameters; segmentation tables were then summarized over genes (CNV-G). All data tables were filtered for downstream analysis by removing features with zero or near-zero variance using the *nearZeroVar* function in the *caret* R package with default parameters. To avoid information leakage, feature filtering was performed on TR sets and applied on both TR and TS sets.

Feature sets. We considered multiple feature sets, namely: TOT, containing the filtered set of features for the corresponding technology (e.g., AG1, MAV, CNV); DEC, a subset of features obtained performing *limma* [15] differential expression analysis on MAV-G TR data between patients with favorable and unfavorable outcome (CLASS endpoint); LIT, a set of 102 genes involved in neuroblastoma retrieved on 2017/05/11 from the literature aggregator Cancer Genetics Web[1].

Predictive classification. We adopted the Data Analysis Protocol (DAP) developed within the MAQC-II and SEQC challenges [16, 17], the U.S. FDA initiatives for reproducibility in microarray and sequencing expression experiments. Briefly, given a dataset split in TR and TS portions, the former undergoes a $10 \times 5-$fold stratified Cross-Validation (CV) resulting in a ranked feature list and a classification performance measure, here the Matthews Correlation Coefficient (*MCC*) [18, 19]. As classifier, we used linear SVMs (with L2L1 penalties) and Random Forest (RF). At each CV iteration, features are ranked by RF Gini index or SVM weights and the classifier is trained on an increasing number of ranked features. A list of top-ranked features is obtained by Borda aggregation of the ranked CV lists [20]. The best model is later retrained on the whole TR set restricted to the features yielding the maximum MCC in CV, and selected for validation on the TS set. As a sanity check to avoid unwanted selection bias effects, the DAP is repeated stochastically scrambling the TR labels (Random Label scheme).

Integrative Network Fusion. INF is a bioinformatics framework for the identification of integrated multi-*omics* biomarkers based on predictive profiling and a novel approach to their integration (Figure 1). In summary, first a RF classifier is trained on the juxtaposed dataset (*juxt*), obtaining a feature list ranked by mean decrease in Gini impurity. Secondly, the data sets are integrated by Similarity Network Fusion [21] and features are ranked by a novel ranking scheme (*rSNF*) based on SNF-fused network clustering; a RF model is developed on the juxtaposed dataset with the feature ranking defined by *rSNF*. From both approaches, a subset of top discriminant features can be identified,

[1] `http://www.cancerindex.org/geneweb/X1701.htm`

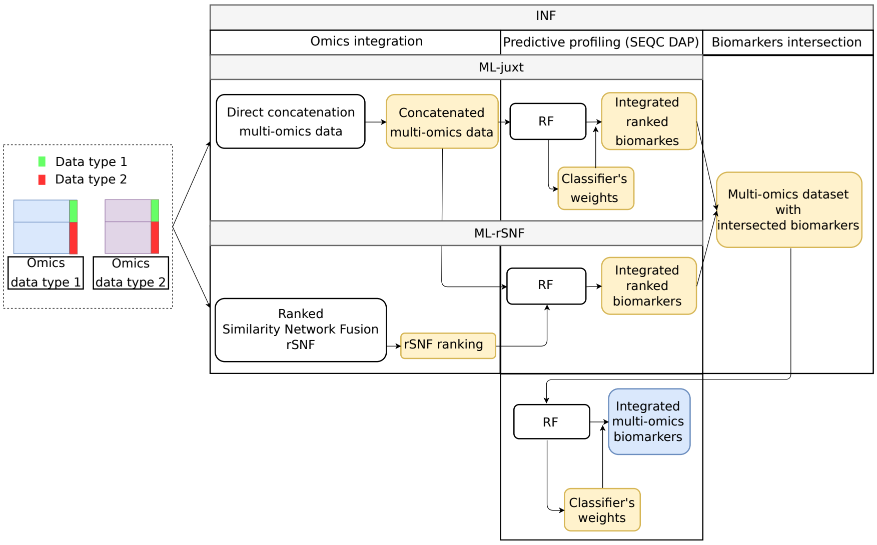
INF
Omics integration
Predictive profiling (SEQC DAP)
Biomarkers intersection
ML-juxt
Direct concatenation multi-omics data
Concatenated multi-omics data
RF
Integrated ranked biomarkes
Classifier's weights
🟩 Data type 1
🟥 Data type 2
Omics data type 1
Omics data type 2
Multi-omics dataset with intersected biomarkers
ML-rSNF
RF
Integrated ranked biomarkers
Ranked Similarity Network Fusion rSNF
rSNF ranking
RF
Integrated multi-omics biomarkers
Classifier's weights

Figure 1: Graphical representation of the INF workflow for two generic *omics* datasets (adapted from [9]). A first RF classifier is trained on the juxtaposed data and the feature list obtained is ranked by mean decrease in Gini impurity (ML-*juxt*). The two data sets are then integrated by Similarity Network Fusion, the features are ranked by *rSNF* and a RF model is developed on the juxtaposed dataset with the feature ranking so defined (ML-*rSNF*). Finally, a RF classifier is trained on the juxtaposed dataset restricted to the intersection of *juxt* and *rSNF* top discriminant feature lists. All the predictive models are developed within the DAP described in the methods.

according to the predictive performance of the classifier. Finally, a RF classifier is trained on the juxtaposed dataset restricted to the intersection of *juxt* and *rSNF* feature lists (*rSNFi*). Predictive models are developed inside the DAP described above. The code implementing INF is available as a GitHub repository[2] (manuscript in preparation).

3. Results

We first applied RF and SVM classifiers to all the sets of data independently, using as labels the endpoints originally proposed in [8] and summarized in Table 1. In general, RF achieved better performance than SVM and predictions based on RNA-Seq were better than those obtained with other techniques. The integration of AG1 with MAV only slightly improved the performance on LIT and DEC feature subsets. Consistently with previous published results, we observed a poor *MCC* performance on HR endpoints. In contrast, the best results were obtained for the CLASS label, identifying patients with extremely positive or negative disease outcomes. SEX label and random labels were used as positive and negative controls respectively. Since only 145 samples were provided with matching aCGH data, we performed similar preliminary analyses also on this subset obtaining comparable results, with CLASS being the best performing endpoint but with generally lower performance likely due to the reduced number of samples available. For this subset we did not consider the HR-OS and HR-EFS endpoints, as the number of samples is too low to allow accurate prediction. Predictions based on CNV alone were generally poor (as shown for ALL-EFS endpoint in the example proposed in Figure 2), while AG1 and MAV performed better and comparably between them. For the integration approach we observed a heterogeneous behavior, with generally better performance on the combined datasets. In particular, as previously found for a meta-omics application in [9], the advantage of INF over simple juxtaposition is a more compact feature signature at similar *MCC* scores.

We present more details on one specific example, namely AG1-G plus CNV-G on the ALL-EFS endpoint, in which the integration gives positive results in terms of both performance on the integrated datasets and compactness of the feature sets retrieved (see Figure 2). All the panels of Figure 2 clearly show that predictions based purely on CNV perform poorly. We observe that with INF we achieve better *MCC* on both CV and TS (Figure 2, panels a and

[2] `https://github.com/AleZandona/INF`

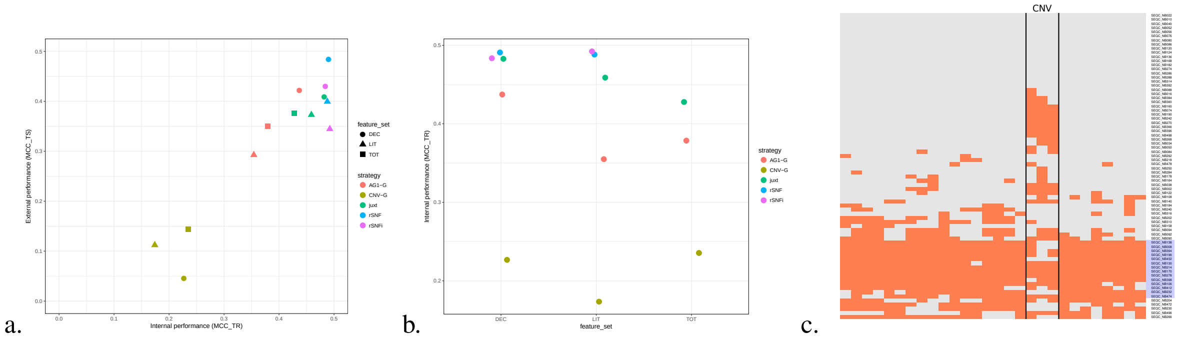

Figure 2: Classification performance on ALL-EFS endpoints for the 145 samples subset. a. Internal (in CV) and external (on TS) performance. b. TS performance stratified by feature set. c. TS predictions heatmap (rows =samples, columns=integration strategy, including unintegrated data; orange: misclassified, grey: correctly predicted). Highlighted in blue are TS samples consistently misclassified.

b). INF identified 75 top features, as opposed to the 4640 identified by simple juxtaposition and the 2171 and 100 for the single datasets (AG1-G and CNV-G, respectively). These include potentially interesting genes. For example, the most important feature extracted by INF is *STK31*, a cell-cycle regulated protein previously implicated in colorectal cancer and recently suggested as novel oncogene [22]. Interestingly, we observe that a set of 12/74 TS samples are consistently misclassified, independent of integration strategy, assay, clinico-genetic subgroups and INSS staging (Figure 2 panel c).

4. Discussion and future directions

We introduced here the INF framework and applied it to the neuroblastoma data made available for the CAMDA2017 challenge. We provide a general overview of the classification performances we obtained across multiple input sets, feature sets and endpoints. We also present details for a case study in which integration allowed us to achieve better performance and more compact feature sets. It is important to notice that our results are part of an ongoing effort aiming at further improving both the technical aspects of the INF implementation and either the performance or the biological insight on this dataset. In particular, our effort is focusing on the following aspects.

1. The CNV calls were shown to perform poorly as classifiers. We hypothesize that improved CNV calls on the aCGH might turn them into a more informative resource. It is important to notice that despite the poor classification performance of CNV alone, their integration with microarray data provided improved classification (Figure 2).
2. Given the promise shown by emerging deep learning approaches applied to biological data and previous successful work from our group in this direction [23], we aim at developing an *omics*-integration framework based on deep learning and applying it to these data.
3. A current limitation of the INF approach for wide feature data resides in the time consuming implementations of the algorithm section performing spectral clustering. A code optimization is currently ongoing, but this limitation currently makes it impossible for us to run INF on the TOT feature set, *i.e.*, including all possible features.
4. Further explore the integration of additional *omics* layers, such as epigenomics data.
5. Interestingly, we identified a set of 12 out of 74 TS samples that are consistently misclassified, independently from integration strategy, assay, clinico-genetic subgroups and INSS staging. This opens the intriguing possibility that these patients could represent a subgroup that may be characterized by distinctive biomarkers.

References

[1] E. Ward, C. DeSantis, A. Robbins, et al., Childhood and adolescent cancer statistics, 2014, CA: A Cancer Journal for Clinicians 64 (2) (2014) 83–103.

[2] E. Newman, J. Nuchtern, Recent biologic and genetic advances in neuroblastoma: Implications for diagnostic, risk stratification, and treatment strategies, Seminars in Pediatric Surgery 25 (5) (2016) 257–264.

[3] M. Esposito, S. Aveic, A. Seydel, et al., Neuroblastoma treatment in the post-genomic era, Journal of biomedical science 24 (1) (2017) 14.

[4] G. Tonini, A. Nakagawara, F. Berthold, Towards a turning point of neuroblastoma therapy, Cancer Letters 326 (2) (2012) 128–134.

[5] B. Salazar, E. Balczewski, C. Ung, et al., Neuroblastoma, a paradigm for big data science in pediatric oncology, International Journal of Molecular Sciences 18 (1) (2016) 37.

[6] S. Riccadonna, G. Jurman, S. Merler, S. Paoli, A. Quattrone, C. Furlanello, Supervised classification of combined copy number and gene expression data, Journal of Integrative Bioinformatics 4 (3) (2007) 74.

[7] M. Wolf, M. Korja, R. Karhu, et al., Array-based gene expression, CGH and tissue data defines a 12q24 gain in neuroblastic tumors with prognostic implication, BMC Cancer 10 (1) (2010) 81.

[8] W. Zhang, Y. Yu, F. Hertwig, et al., Comparison of RNA-seq and microarray-based models for clinical endpoint prediction, Genome Biology 16 (1) (2015) 133.

[9] A. Zandonà, Predictive networks for multi-omics data integration, PhD thesis, Centre for Integrative Biology, University of Trento, Italy (2017).

[10] S. Stigliani, S. Coco, S. Moretti, et al., High genomic instability predicts survival in metastatic high-risk neuroblastoma, Neoplasia 14 (9) (2012) 823IN6–832IN10.

[11] S. Coco, J. Theissen, P. Scaruffi, et al., Age-dependent accumulation of genomic aberrations and deregulation of cell cycle and telomerase genes in metastatic neuroblastoma, International journal of cancer 131 (7) (2012) 1591–1600.

[12] H. Kocak, S. Ackermann, B. Hero, et al., Hox-c9 activates the intrinsic pathway of apoptosis and is associated with spontaneous regression in neuroblastoma, Cell death & disease 4 (4) (2013) e586.

[13] J. Theissen, A. Oberthuer, A. Hombach, et al., Chromosome 17/17q gain and unaltered profiles in high resolution array-cgh are prognostically informative in neuroblastoma, Genes, Chromosomes and Cancer 53 (8) (2014) 639–649.

[14] F. Commo, J. Guinney, C. Ferté, et al., rcgh: a comprehensive array-based genomic profile platform for precision medicine, Bioinformatics 32 (9) (2016) 1402.

[15] G. Smyth, Limma: linear models for microarray data, in: Bioinformatics and computational biology solutions using R and Bioconductor, Springer, New York, 2005, pp. 397–420.

[16] The MicroArray Quality Control (MAQC) Consortium, The MAQC-II Project: A comprehensive study of common practices for the development and validation of microarray-based predictive models, Nature Biotechnology 28 (8) (2010) 827–838.

[17] The SEQC/MAQC-III Consortium, A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequence Quality Control consortium, Nature Biotechnology 32 (2014) 903–914.

[18] B. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochimica et Biophysica Acta 405 (2) (1975) 442–451.

[19] P. Baldi, S. Brunak, Y. Chauvin, et al., Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (5) (2000) 412–424.

[20] G. Jurman, et al., Algebraic stability indicators for ranked lists in molecular profiling, Bioinformatics 24 (2) (2008) 258–264.

[21] B. Wang, A. M. Mezlini, F. Demir, et al., Similarity network fusion for aggregating data types on a genomic scale, Nature methods 11 (3) (2014) 333–337.

[22] A. Schlicker, M. Michaut, R. Rahman, L. Wessels, OncoScape: Exploring the cancer aberration landscape by genomic data fusion, Scientific Reports 6 (1) (2016) 28103.

[23] C. Zarbo, V. Maggio, M. Chierici, et al., Integrating deep learning with the SEQC data analysis plan for predictive biomarkers of clinical endpoints in neuroblastoma, DLPM2016 @ ECML-PKDD2016 (2016).