A multi-layer network approach to data integration for patient stratification

Maciej M Kańduła,[1] Swati Singh,[1,2] Eric D Kolaczyk[3], and David P Kreil[1]

[1]Chair of Bioinformatics Research Group, Dept of Biotechnology, BOKU University Vienna, Austria

[2]Dept of Biological Sciences and Bioengineering, Indian Institute of Technology, Kanpur, India

[3]Dept of Mathematics and Statistics, Boston University, USA

contact: maciej.kandula@boku.ac.at; correspondence: kolaczyk@bu.edu, david.kreil@boku.ac.at

In recent years we have been witnessing a great increase of available biomedical data that is being collected from high-throughput experiments, with many datasets of genomic scale (Benton et al, 1996; Mushegian et al, 2011). The Big Data bottleneck experienced for both basic and applied research, however, which remains rate-limiting in the translation of experimental advances to the clinic, is the identification and interpretation of biologically relevant patterns in these data. A lot of hope is now being placed in analyses approaches which combine measurements from these different sources (Searls et al, 2005). In order to further explore and take advantage of complex big data sets, novel and increasingly sophisticated computational methods are being developed. Different data types may capture complementary aspects of information and, investigated together, could help shed light on complex relationships of interest, such as between the impact of gains and losses in gene copies and the cis or trans expression of specific genes related to tumour progression, and the eventual consequences for cancer survival time prediction. Indeed, a variety of analyses incorporating multiple sources of evidence have been reported to help advance our understanding of biological systems (Hartemink et al, 2002; Nariai et al, 2005; Hecker et al, 2009). With complementary data collected for each patient, it is natural to try and combine all a patient's interrelated information into a single joint representation. Recently, network representations have been explored to exploit not just the complementary nature of the data sources but also similarities across patients (Wang et al, 2014).

Networks have already been applied to identify dysregulated pathways(Pham et al, 2011; Verbeke et al, 2015), optimize biotechnological processes (Mizrachi et al, 2017), and predict patient survival (Wang et al, 2014). We here introduce a novel network-based approach for the integration of multiple molecular and clinical data types that also incorporate prior knowledge from curated databases, exploring performance in comparative quantitative benchmarks. Our algorithm creates a multiple-layer network (Kivelä et al, 2014), where the highest level is a network of patients, and each patient has information from multiple data types, where each data type is characterized itself by a network (Fig.1). Notably, prior functional knowledge is incorporated in their construction. This structure facilities an

identification of similarities not only between patients but also of functional modules at the molecular level, across data-types.

TCGA datasets were studied in previous CAMDA challenges and remain a valuable source of data for large-scale analyses. In this extended abstract, we report first results of applying our network analysis to TCGA data sets for a well studied cancer, Breast Invasive Carcinoma (BRCA). These data are particularly well suited for initial method validation, as it comprises a large cohort of patients with matched profiles for genomic methylation, RNA-Seq and miRNA-Seq gene expression, complemented by detailed clinical information. Our work will be extended to also examine this year's CAMDA neuroblastoma data sets exploiting dual gene expression and matched copy-number data in time for the conference in July.

In order to examine the explanatory power of our network we first develop a personalized approach to patient stratification. We will show that already this first application of our algorithm considerably improves on a related state-of-the-art network approach (Wang et al, 2014) in cancer patient stratification, showing a significant prognostic effects in Cox regression and Hazard Ratio (HR) of >30% already from the stratification alone. Subsequent work will characterize the underlying molecular modules and their clinical relevance.

Method description

We link patients in a multi-layer network with *N* elementary networks per layer, where a node represents a patient (Fig. 1) and lower-level networks are created for each measurement type provided for the patient (Fig. 2). The measurements can be, *e.g.*, gene expression, copy number experiments, methylation of genomic DNA, *etc*.

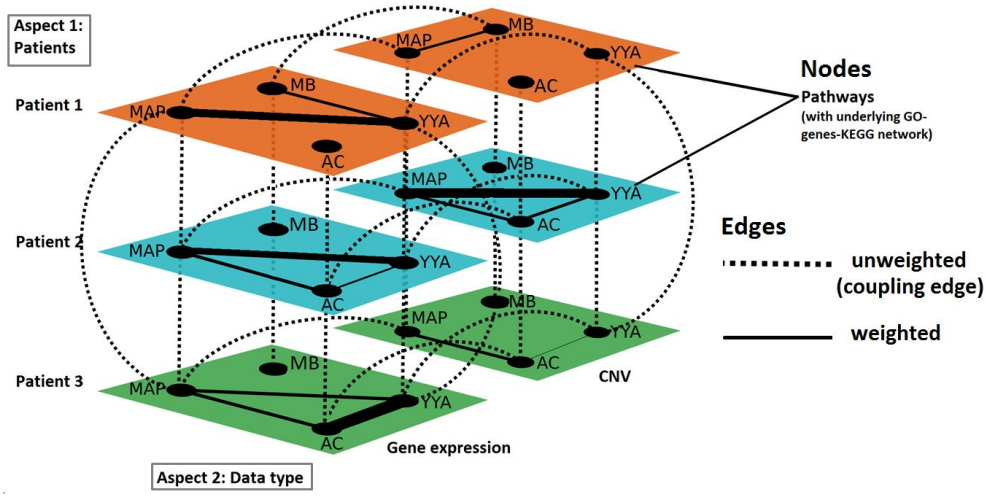

Figure 1: multi-layer network design

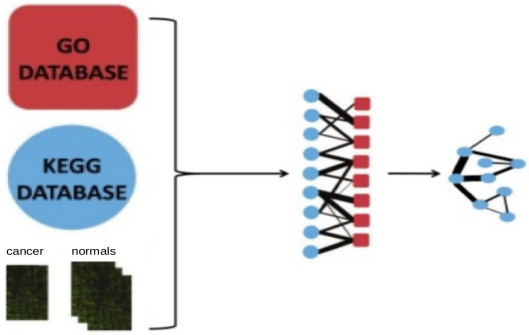

Figure 2: elementary layer design

In this setting we are looking for similarities (and possibly differences) between the nodes, *i.e.*, patients. In general, for any choice of a patient-specific multi network, for each patient pair, we define a level-by-level metric and stack those metrics into a vector. In the patient network we thus obtain vector-valued edge weights (Fig. 3), reflecting similarity across different layers of biology that define our multi networks.

For constructing the lower-level 'elementary' networks (Fig. 2) we apply a novel data integration approach based on a recently published algorithm (Pham et al, 2011). It uniquely combines measurements and existing knowledge by integrating structured information from several sources, collecting genes into an interconnected network of functionally-related gene sets, through KEGG pathways and GeneOntology functional groups. Edges between nodes are weighted according to a probabilistic strength of evidence for differential effect in genes relevant to the biological functions in which the corresponding biological pathways are involved. In the original application this network was applied to identification of dysregulated pathways in cancer. We here use the network of KEGG gene sets as the first layer in our multi-layered patient network.

Research on multiple-layer networks is novel and ongoing, and there are no established algorithms for meaningful clustering within our complex network that could readily be applied. We therefore start by looking for clusters of similar gene sets in each elementary network of KEGG gene sets of a patient. We then tested spectral clustering instead for this step (Ng et al, 2001; Luxburg, 2007) but resulting groups were of no obvious value for prognosis. We could, however, successfully exploit stochastic block models (Leger, 2014) in identifying meaningful clusters. Stochastic block models (SBM) were previously applied to multi-layer networks with only one elementary network *per* entity (Stanley et al, 2016). Here, we successfully generalize this step-wise to multi-layer networks accommodating more than one data-type, where we apply the algorithm to the different data types one at a time. Integrated Classification Likelihood (Biernacki et al, 1998) was used for selecting the optimal number of clusterings for each network. We have since also considered searching for recurrent heavy subgraphs (Li et al, 2011) in our network, and research in this direction is still ongoing.

We can then compare, across patients, the KEGG gene set groups identified by SBM for each specific data type by using a shared information distance between sets of clusterings, namely

the variation of information (VI) metric (Meilă, 2007). Its characteristics as a metric ensure that patient-patient distances can directly be compared between the networks constructed for the different data types. This eventually yields vector-valued edge weights between patients, with each vector coordinate corresponding to a data type (Fig 3. right-hand network).

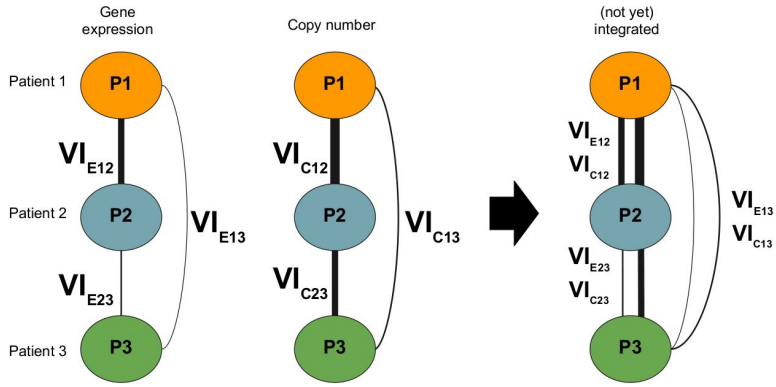

Figure 3: patient-patient network

Finally, these vectors can be summarized by computing the average VI metric across data types for each edge. We can then successfully identify similar patients using a spectral clustering algorithm (Ng et al, 2001; Luxburg, 2007), analogous to earlier network analyses (Wang et al, 2014). Here the optimal number of clusters was selected by gradient descent (Zelnik-Manor & Perona, 2005).

Results and Discussion

In order to evaluate if our constructed network captures meaningful patient-patient relationships, we performed Cox proportional hazards regression analysis on the patient stratification arising from the network. We compare the predictive power of the groups of patients obtained by our network approach with the groups identified by SNF, an alternative state-of-the-art network-based method (Wang et al, 2014). In addition we examine the effects of data integration in our approach, comparing results for integrative and single-track analyses. Cox log-rank test helps in evaluating the significance of differences in survival profiles between subgroups. The Hazard Ratio (HR) further explains how big an impact being in a particular patient group has on survival. As shown in Tab. 1 our integrative multi-layer network approach finds highly significant differences in survival among patient groups and with a substantial effect size (>30%). Apart from ‘age’ alone, which has only a very small effect on the survival time of a patient, none of the other examined predictions was significant, including the single-track or integrative network-based analysis by SNF, showing the value of our integrative network for exploring a patient-patient relationships.

method / variable	HR	p-val (Log-rank test)
age	1.04	4.4×10^{-4}
multi-layer	0.69	4.3×10^{-3}
age+'multi-layer'	0.73	1.5×10^{-4}
RNA-Seq	1.04	0.59
mirSeq	1.94	0.08
methylation	0.83	0.34
SNF (5 clusters)	1.28	0.15
age+SNF (5 clusters)	1.2649	9.6×10^{-4}

Table 1: comparison of predictive power of patient grouping using different methods / variables

With these encouraging initial we will apply our approach to other cancers and data types, specifically the CAMDA 2017 neuroblastoma dataset, which provides RNA-Seq and microarray expression, genomic copy number data, and matched clinical profiles.

To improve the immediate clinical relevance of our work, we will examine the performance of predicting the survival time of a single hitherto unseen patient. Here the network will first be build with a representative cohort of patients, creating a patient-patient multi-layer network for context. Per-patient networks of each data type and across data types can then be built for any 'new' patients. These are then merged into the patient-patient network, facilitating prognosis by the respective patient-patient network neighbourhood. The charm of this approach is that any additional patients improve on the original network in partially supervised predictions. As more diagnostic data becomes available for new patients, the additional labels immediately improve prediction power.

We will also examine the relevance of the functional modules identified in lower-level networks. In particular, we will examine network-based Cox regression, which has recently been applied to identify biomarkers predictive of survival times by using gene co-expression networks to inform sparse Cox regression (Iuliano et al, 2016). We can directly apply this idea to our elementary networks of KEGG pathways *per* data type, testing if the additional abstraction level to functional modules and the incorporation of prior knowledge through the GeneOntology used in the construction of these networks helps identify strong predictors at the level of functional modules. It will particularly be interesting to test if this improves generalization accuracy across different (sub)cohorts, a challenge often underestimated in the identification of (potentially confounded) biomarkers from single cohort studies. The aim would be to find the most relevant biomarkers in a cohort that will generalize well to future patients, rather than the incidentally strongest in a single cohort. Future work can then investigate non-uniform network-based weighting of patients in this context, *i.e.*, making sure all patient types enter the regression with comparable weights. This may be particularly relevant for seeking functional insights.

In future work (after the conference) we will also compare alternative measures of multi-layer network similarity, *e.g*, exploring heavy subgraph detection.