PATIENT-SPECIFIC DATA FUSION FOR CANCER STRATIFICATION AND PERSONALISED TREATMENT

VLADIMIR GLIGORIJEVIĆ, NOËL MALOD-DOGNIN AND NATAŠA PRŽULJ*

Department of Computing, Imperial College London,
London, SW7 2AZ, United Kingdom
**E-mail: natasha@imperial.ac.uk*
http://www.doc.ic.ac.uk/~natasha

According to Cancer Research UK, cancer is a leading cause of death accounting for more than one in four of all deaths in 2011. The recent advances in experimental technologies in cancer research have resulted in the accumulation of large amounts of patient-specific datasets, which provide complementary information on the same cancer type. We introduce a versatile data fusion (integration) framework that can effectively integrate somatic mutation data, molecular interactions and drug chemical data to address three key challenges in cancer research: stratification of patients into groups having different clinical outcomes, prediction of driver genes whose mutations trigger the onset and development of cancers, and repurposing of drugs treating particular cancer patient groups. Our new framework is based on graph-regularised non-negative matrix tri-factorization, a machine learning technique for co-clustering heterogeneous datasets. We apply our framework on ovarian cancer data to simultaneously cluster patients, genes and drugs by utilising *all* datasets. We demonstrate superior performance of our method over the state-of-the-art method, Network-based Stratification, in identifying three patient subgroups that have significant differences in survival outcomes and that are in good agreement with other clinical data. Also, we identify potential new driver genes that we obtain by analysing the gene clusters enriched in known drivers of ovarian cancer progression. We validated the top scoring genes identified as new drivers through database search and biomedical literature curation. Finally, we identify potential candidate drugs for repurposing that could be used in treatment of the identified patient subgroups by targeting their mutated gene products. We validated a large percentage of our drug-target predictions by using other databases and through literature curation.

Keywords: Data fusion; Tumor stratification; Drug repurposing; Cancer driver genes; Non-negative Matrix Factorization.

1. Introduction

Cancer is a leading cause of morbidity worldwide. It is a complex genetic disease in which the genomes of normal cells accumulate somatic mutations and other alterations that are eventually perturbing vital cellular functions. Recent advances in DNA sequencing technologies have enabled identification of somatic mutations across tumor genomes and exomes of individual patients[1,2]. These somatic mutations provide a new and rich source of data for addressing many challenges in cancer research, such as indentifying driver genes (i.e., genes whose mutations lead progression of oncogenesis), stratifying patients into biologically meaningful classes with different clinical outcomes and creating new opportunities for development of successful personalized treatment strategies[3]. Cancer is also a highly heterogeneous disease with large genetic diversity even between tumors of the same cancer type. Namely, two clinically identical tumors rarely have a large set of common mutated genes. Moreover, very few genes are frequently mutated across tumor samples. This makes the use of somatic mutations for iden-

tification of driver genes, as well as for patient stratification into subtypes, much harder[1,4,5]. However, despite this genetic diversity between tumor samples, the perturbed pathways are often similar[1]. Therefore, integration of somatic mutations with other genomic data, such as with molecular networks that contain pathways, is a promising direction for addressing these problems.

Development of computational methodologies that can efficiently integrate genome-scale molecular information and address the above mentioned challenges in cancer research is of foremost importance. A majority of previous studies do not utilise data on somatic mutations, but instead, they are mainly based on mRNA expression and methylation data. Because of noisiness of these data, the patient stratification studies for cancer types often do not produce patient subgroups that agree well with any clinical or survival data[6]. Therefore, a recent study proposed the use of somatic mutation data in combination with biological networks as a new source of information for tumor stratification[5]. However, the proposed methodology cannot account for additional data types (e.g., drug data) and cannot be used for identifying novel driver genes, nor for predicting a personalised therapy. Moreover, previous data integration methods can only be used for either cancer patient stratification[5], driver gene prediction[7] or drug repurposing[8].

Here, we present a versatile patient-specific data integration (fusion) methodology capable of: 1) uncovering patient subgroups (stratification) with prognostic survival outcome, 2) predicting novel driver genes and 3) repurposing drugs, i.e., predicting new candidate drugs for targeting mutated gene products in individual patients and that can be used in treatment of identified patient subgroups. To our knowledge, this is the first method that can address all three challenges simultaneously. Our methodology is based on Non-negative Matrix Tri-Factorization (NMTF) technique, initially proposed for dimensionality reduction and co-clustering problems in machine learning[9]. It approximates (factorises) a high-dimensional data matrix, representing relations between two data types, as a product of three non-negative, low-dimensional matrices[9]. The clustering interpretation of low-dimensional matrices and their previously established relatedness to the k-means clustering has enabled the use of NMTF in co-clustering problems[10,11]. Recently, there has been a significant development in the use of NMTF in data fusion because of its ability to extend to any number of interrelated data types by *simultaneously* decomposing their relation matrices. This has provided us with a valuable framework for fusion (integration) of any number and type of interrelated heterogeneous datasets[12,13]. NMTF has demonstrated a great potential in addressing various biological problems, such as disease association prediction[12], disease gene discovery[14], protein-protein interaction prediction[15] and gene function prediction[16].

We use NMTF to integrate somatic mutation profile (SMP) data of serous ovarian cancer patients from TCGA[4] with molecular networks (MNs) from BioGRID[17] and KEGG[18], drug-target interaction (DTI) and drug chemical similarity (DCS) data from DrugBank[19] (detailed in Sec. 2.3). We perform consensus clustering by using NMTF to simultaneously cluster patients, genes and drugs based on the evidence from *all* datasets. First, we stratify patients into three groups that we assess by using clinical data. We show significant difference in survival outcomes between these groups, as well as a good agreement with other clinical data.

Second, from clusters of genes, we identify those enriched in known driver mutations; we postulate genes strongly related to known driver genes in these clusters as potential drivers genes, i.e., genes responsible for ovarian cancer progression. Finally, we use the matrix completion property of NMTF to predict new drug-target relations and to identify new drug candidates that could be used for repurposing and treatment of identified ovarian cancer patient groups. Furthermore, we evaluate the influence of all combinations of datasets onto the accuracy of drug-target predictions by performing a 5-fold cross validation. We shown that the highest accuracy is achieved when all datasets are taken into account, proving the utility of integrating all considered datasets (detailed in Sec. 3).

2. Methods

2.1. *Patient-specific data fusion framework*

We consider there different datasets: patients, genes and drugs. Patients and genes are related to each other by somatic mutation profiles (SMPs), constructed for n_1 patients over n_2 genes and encoded in high-dimensional relation matrix, $\mathbf{R}_{12}^{n_1 \times n_2}$. Its entries are binary values, with $\mathbf{R}_{12}[p][g] = 1$ if gene g is found to be mutated in patient p, and zero otherwise. Genes and drugs are related to each other according to drug-target interactions (DTIs). DTIs between n_2 genes and n_3 drugs are encoded in relation matrix, $\mathbf{R}_{23}^{n_2 \times n_3}$. Its entries are also binary values, with $\mathbf{R}_{23}[g][d] = 1$, if the product of gene g is a target of drug d and zero otherwise. See Sec. 2.3 and Fig. 1 for details of construction of the relation matrices and for an illustration of these datasets.

We use NMTF to simultaneously decompose both relation matrices into a product of three non-negative low-dimensional matrices as follows: $\mathbf{R}_{12} \approx \mathbf{G}_1\mathbf{H}_{12}\mathbf{G}_2^T$. and $\mathbf{R}_{23} \approx \mathbf{G}_2\mathbf{H}_{23}\mathbf{G}_3^T$. The low dimensional matrices can be obtained by solving the following optimisation problem: $\min_{\mathbf{G}_i \geq 0, 1 \leq i \leq 3} J = \min_{\mathbf{G}_i \geq 0, 1 \leq i \leq 3} \Big(\parallel \mathbf{R}_{12} - \mathbf{G}_1\mathbf{H}_{12}\mathbf{G}_2^T \parallel_F^2 + \parallel \mathbf{R}_{23} - \mathbf{G}_2\mathbf{H}_{23}\mathbf{G}_3^T \parallel_F^2 \Big)$, where F denotes Frobenius norm and J is the objective function. The non-negativity constraints imposed on $\mathbf{G}_i$ matrices for $1 \leq i \leq 3$ provide easier interpretation of their values in the clustering assignment. Many of the data types are characterised by additional, internal connectivity structure represented by graphs (networks). In this study, genes are connected by molecular networks (MNs), while drugs are connected based on the similarity of their chemical structures, i.e., drug chemical similarity (DCS) network (illustrated in Fig. 1). We incorporate these networks (MN and DSC) into the above objective function by adding two regularisation terms to constrain the construction of $\mathbf{G}_2$ and $\mathbf{G}_3$ matrices. This approach is also known also as *Graph-regularized NMTF* (or GNMTF)[20]. Namely, the aim is to enforce two interacting genes to belong to the same cluster (similarly with drugs) and a violation of these constrains results in penalties to our objective function. Hence, the final objective function has the following form:

$$\min_{\mathbf{G}_i \geq 0, 1 \leq i \leq 3} J = \min_{\mathbf{G}_i \geq 0, 1 \leq i \leq 3} \big[\parallel \mathbf{R}_{12} - \mathbf{G}_1\mathbf{H}_{12}\mathbf{G}_2^T \parallel_F^2 + \parallel \mathbf{R}_{23} - \mathbf{G}_2\mathbf{H}_{23}\mathbf{G}_3^T \parallel_F^2 + tr(\mathbf{G}_2^T\mathbf{L}_2\mathbf{G}_2) + tr(\mathbf{G}_3^T\mathbf{L}_3\mathbf{G}_3) \big] \quad (1)$$

where, tr denotes the trace of a matrix, and $\mathbf{L}_2$ and $\mathbf{L}_3$ are graph Laplacians of MN and DCS networks, respectively.

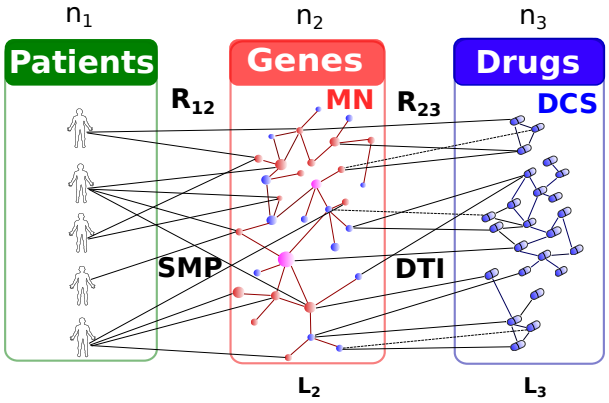

Fig. 1. Schematic illustration of datasets used in this study. Three types of objects (nodes) are represented: n_1 *ovarian cancer patients* (in green), n_2 *genes* (in red) and n_3 *drugs* (in blue). Somatic mutation profiles (SMP) for ovarian cancer patients are represented by patient-gene links denoting assignment of mutated genes (in red) to each individual patient. These connections are encoded into relation matrix, $\mathbf{R}_{12}$. Genes are connected by a molecular network (MN) obtained by merging three different interaction data types (see Sec. 2.3). Also, MN is the union of three types of genes: *mutated genes* coming from patients' SMPs (in red), *driver genes* (in pink) coming from TCGA database and *normal genes* (in blue) coming from other databases used for construction of networks (details in Sec. 2.3). Connections in this network, MN, are represented by Laplacian matrix, $\mathbf{L}_2$. Links between genes (i.e., their protein products, that are drug targets) and drugs are drug-target interactions (DTIs) and are represented by relation matrix, $\mathbf{R}_{23}$. Links between drugs are represented by drug chemical similarity (DCS) network (details in Sec. 2.3). Connections in this network, DCS, are represented by Laplacian matrix, $\mathbf{L}_3$.

The key idea of our GNMTF-based data fusion approach is in sharing low-dimensional matrix $\mathbf{G}_2$ whilst simultaneously learning from (i.e., decomposing) relation matrices, $\mathbf{R}_{12}$ and $\mathbf{R}_{23}$. Such decomposition accounts for collective influence of all data sets (along with molecular and chemical constraints effectively integrated within the same framework) onto the resulting clustering of patients, genes and drugs. This approach corresponds to the *intermediate* data fusion in which the structure of the data is preserved during the model inference. Such an approach has been shown to result in the best accuracy among all data fusion approaches[12].

Minimisation of the objective function, J, is done by *multiplicative update rules* used to compute all low-dimensional matrices; under these multiplicative rules, the objective function is non-increasing[11]. The minimisation starts by randomly initialising $\mathbf{G}_i$ matrices for $1 \leq i \leq 3$ and then iteratively updating their values until the convergence criterion is reached. In all our runs, we use *Random Acol* initialisation strategy[21] and the convergence criterion is reached when $\frac{|J_{n+1}-J_n|}{|J_n|} < 10^{-5}$. The multiplicative update rules, their derivation and proof of convergence can be found in Wang *et al.*[11].

Co-clustering of patients, genes and drugs. Matrices $\mathbf{G}_1^{n_1 \times k_1}$, $\mathbf{G}_2^{n_2 \times k_2}$ and $\mathbf{G}_3^{n_3 \times k_3}$ from Equation 1 above are *cluster membership indicator* matrices for patients, genes and drugs, respectively; based on their entries, n_1 patients are assigned to k_1 patient clusters, n_2 genes are assigned to k_2 gene clusters and n_3 drugs are assigned to k_3 drug clusters, respectively. In particular, following the *hard clustering* procedure of Brunet *et al.*[22], matrix $\mathbf{G}_1^{n_1 \times k_1}$, with rows representing patients and columns representing clusters, is used to place patient p into

cluster k if $\mathbf{G}_1[p][k]$ is the largest entry in row p. This assignment procedure results in the binary connectivity matrix for patients, $\mathbf{C}_1^{n_1 \times n_1}$, with entry $\mathbf{C}_1[p_1][p_2] = 1$ if patients p_1 and p_2 belong to the same cluster and $\mathbf{C}_1[p_1][p_2] = 0$ otherwise. We apply this procedure for all cluster membership indicator matrices. The number of clusters (also called *rank parameters*) for each dataset are chosen to be $k_1 \ll n_1$, $k_2 \ll n_2$ and $k_3 \ll n_3$, which provides dimensionality reduction of the relation matrices. Matrices $\mathbf{H}_{12}^{k_1 \times k_2}$ and $\mathbf{H}_{23}^{k_2 \times k_3}$ in Equation 1 above represent compressed, low-dimensional versions of $\mathbf{R}_{12}$ and $\mathbf{R}_{23}$, respectively.

An important step in our methodology is estimating rank parameters, which are the numbers of clusters of patients, genes and drugs, k_1, k_2 and k_3, respectively. These parameters need to be known before factorisation is performed. The usual procedure for obtaining these parameters is by varying these parameters for each run and estimating cluster stability[22,23]. We take the values of parameters for which the most stable clustering is achieved. In particular, multiplicative update rules converge to a different solution in each run, depending on the random matrix initializations (i.e., initial clustering assignment given by the initial values in matrices $\mathbf{G}_i$, $1 \leq i \leq 3$). For example, if a clustering of patients into k_1 classes is stable, we expect small variations in the assignment to clusters from run to run. To measure this, we perform multiple factorisation runs with the same values of rank parameters. Each time, a connectivity matrix is computed (e.g., $\mathbf{C}_1$ for patients); based on these, an averaged connectivity matrix (also called the *consensus* matrix) over all runs is computed, $\hat{\mathbf{C}}_1$. If the clustering is stable, then the entries in $\mathbf{C}_1$ (also referred to as the *cluster association scores*) will be either close to zero, or close to one. Otherwise, the entries will be scattered in the interval $[0, 1]$. We use the *dispersion* coefficient, $\rho_{k_1}(\hat{\mathbf{C}}_1)$, introduced by Kim *et al.*[23], as a measure of cluster stability. The values of the dispersion coefficient range in $0 \leq \rho_{k_1}(\hat{\mathbf{C}}_1) \leq 1$, where 1 denotes a stable clustering. In our study, for each rank parameter, we perform a grid search in intervals of 1 for $1 \leq k_1 \leq 5$, $5 \leq k_2 \leq 30$ and $5 \leq k_3 \leq 30$, and compute dispersion coefficients, $\rho_{k_1}(\hat{\mathbf{C}}_1)$, $\rho_{k_2}(\hat{\mathbf{C}}_2)$ and $\rho_{k_3}(\hat{\mathbf{C}}_3)$ for patients, genes and drugs, respectively. We choose the values for k_1, k_2 and k_3 for which dispersion coefficients are of the highest values.

Matrix completion property. In addition to co-clustering of patients, genes and drugs, we model the existing and predict new drug-target interactions by using the *matrix completion* property of GNMTF. Namely, after obtaining low-dimensional matrices, the reconstructed drug-target matrix, $\hat{\mathbf{R}}_{23} \approx \mathbf{G}_2\mathbf{H}_{23}\mathbf{G}_3^T$, is more complete than the initial matrix, $\mathbf{R}_{23}$, and it can be used for extracting new, unobserved drug-target relations and therefore, finding new drug candidates for repurposing.

2.2. *Drug repurposing, patient stratification and driver gene prediction*

Drug repurposing. We use the reconstructed drug-target relation matrix, $\hat{\mathbf{R}}_{23}$, to extract new, previously, unobserved drug-gene interactions and to postulate new candidates for drug repurposing in the treatment of ovarian cancer patients. We apply a combination of row-centric and column-centric rules to extract new, strongly associated drug-gene pairs[13]. Namely, a drug-gene pair, (d, g), is considered to be predicted, if the estimated association score, $\hat{\mathbf{R}}_{23}[g][d]$, is greater than the mean association score of all relations of gene g, as well as greater than the

mean association score of all relations of drug d.

Patient stratification. We stratify ovarian cancer patients into groups, according to the consensus matrix, $\hat{\mathbf{C}}_1$. We use the approach of Brunet *et al.*[22]: we use the off-diagonal entries of $\hat{\mathbf{C}}_1$ as a measure of patient similarity and apply average linkage hierarchical clustering to group patients into k_1 classes. Results and validations are shown in Sec. 3.1 below.

Cancer driver gene prediction. Similar to the patient consensus matrix, we use the gene consensus matrix, $\hat{\mathbf{C}}_2$, to extract gene clusters and identify those that are enriched in mutations and known driver genes by using *the standard model sampling without replacement test (i.e., hypergeometric test)*. In clusters that are enriched in known drivers, we identify genes that are highly associated with known driver genes based on the clustering association scores from the gene consensus matrix. We postulate that these genes are new driver genes for ovarian cancer. Results and validations are presented in Sec. 3.2 below.

2.3. *Datasets, pre-processing and matrix construction*

We downloaded high-grade serous ovarian cancer somatic mutation data from TCGA data portal[4] on the 2^{nd} of July 2015. We only consider data generated by using Illumina GAIIx platform, having the largest number of patients. Following the same procedure for data filtering as in Hofree *et al.*[5], we retain only the patients with more that 10 mutated genes. This results in $n_1 = 353$ serous ovarian cancer patients with mutations in the total of 11,148 genes. Mutated genes are mapped onto the Molecular Network (MN) that we obtain by merging three different biological networks: protein-protein interaction (PPI) and genetic interaction (GI) network from BioGRID database (version 3.4.126)[17], and metabolic interaction (MI) network from KEGG database[18]. This results in MN of 236,751 interactions among $n_2 = 19,118$ genes (mutated and normal). We represent these interactions by Laplacian matrix, $\mathbf{L}_2^{n_2 \times n_2}$, computed as: $\mathbf{L}_2 = \mathbf{D}_2 - \mathbf{A}_2$, where $\mathbf{A}_2$ is the adjacency matrix of MN and $\mathbf{D}_2$ is the diagonal degree matrix of MN (i.e., whose entries on the diagonal are row sums of $\mathbf{A}_2$ and all other entries in $\mathbf{D}_2$ are zeros). For each patient, we create an n_2- long binary $(0, 1)$ somatic mutation profile (SMP) vector, where "1" indicates the existence of a mutated gene in the patient and all other entries are "0". These mutation profiles for all n_1 patients are captured in a binary relation matrix $\mathbf{R}_{12}^{n_1 \times n_2}$ consisting of these SMP vectors. Due to the sparsity of matrix $\mathbf{R}_{12}$, we apply a network propagation technique as the pre-processing step to smooth the patient profiles, by spreading the influence of each mutation over its neighbours in MN network. We use the network propagation method proposed by Vanunu *et al.*[24], based on which the new patient-gene relation matrix is computed iteratively as follows: $\mathbf{R}_{12}^{t+1} = \alpha \mathbf{R}_{12}^{t} \bar{\mathbf{A}}_2 + (1-\alpha)\mathbf{R}_{12}^{0}$, where $\bar{\mathbf{A}}_2$ is the normalised adjacency matrix of MN network computed as $\bar{\mathbf{A}}_2 = \mathbf{A}_2 \mathbf{D}_2^{-1}$, $\mathbf{R}_{12}^{0} = \mathbf{R}_{12}$ is the initial patient-gene matrix and α is a tuning parameter that controls the distance of diffusion through MN network. In all our runs, we set $\alpha = 0.6$ (as it produced the best results), and we took the final network-smoothed, patient-gene matrix (after convergence, $|\mathbf{R}_{12}^{t+1} - \mathbf{R}_{12}^{t}| < 10^{-6}$, is achieved) as input to GNMTF. This pre-processing step has been shown to lead to much better and more robust patient stratification results in previous studies[5], hence we use it as well.

Drug-target interactions (DTIs) are downloaded from DrugBank database (version 4.3)[19]. We retrieved $n_3 = 6,620$ drugs (FDA-approved and experimental) targeting 1,385 genes in MN. These interactions are captured by DTI binary relation matrix, $\mathbf{R}_{23}^{n_2 \times n_3}$. SMILES chemical representation of the n_3 drugs are also retrieved from DrugBank database. The two-dimensional chemical similarity between drugs are computed by using Tanimoto similarity coefficient[25]. We retain only the top 5% most similar drug pairs, which results in 1,069,393 links in the drug chemical similarity (DCS) network. We represent these links by Laplacian matrix, $\mathbf{L}_3^{n_3 \times n_3}$ (computed in the same way as for MN network, described above).

2.4. *Clinical and biological validation of results*

For all patients, we also downloaded clinical follow up data from TCGA database, including the overall patients' survival information (days to the last follow-up and vital status), age, tumor grade, size and tumor position. We used these data to assess the clinical relevance of the patient clusters that we obtain after data fusion. We used Kaplan-Meier survival curves, as well as the log-rank p-value, to measure the significance of the difference in survival profiles between different patient clusters. The log-rank p-value measures the probability of the null hypothesis that patients in each cluster are drawn from the same underlying survival distribution[26]. From TCGA database, we also retrieved a list of 83 known ovarian cancer driver genes, out of which 76 are present in our set of mutated patient genes. We use this set of genes to assess gene clusters obtained after fusion and to identify clusters enriched in drivers.

3. Results

3.1. *Patient stratification*

We perform the consensus clustering, as described in Sec. 2.1, with 20 different random initialisations (initial cluster assignment) of GNMTF and compute the consensus matrices of patients, genes and drugs. We observe that rank parameters of $k_1 = 3$ (the number of patient clusters), $k_2 = 25$ (the number of gene clusters) and $k_3 = 20$ (the number of drug clusters), lead to the most stable clustering (with the largest dispersion coefficients: $\rho_{k_1=3}(\hat{\mathbf{C}}_1) = 0.56$, $\rho_{k_2=25}(\hat{\mathbf{C}}_2) = 0.91$ and $\rho_{k_3=20}(\hat{\mathbf{C}}_3) = 0.88$).

To assess the prognostic capabilities of our patient-specific data fusion approach on ovarian cancer patients, we perform clinical validation of the three obtained patient clusters. The Kaplan-Meier survival curves, shown in Fig.2 (A), reveal the low-survival group (*Cluster 2*) with 56% of death cases and the good outcome group (*Cluster 1*) with 38% of death cases. We observe that the identified clusters are highly discriminative with the log-rank p-value of 5.3×10^{-3}. The same number of clusters has been also reported in previous studies done on somatic mutation and molecular interaction data[5], and also in study done only on miRNA expression data[4]. Furthermore, the identified clusters display a good agreement with the median age of patients in clusters, with *Cluster 2* having the oldest patients. In addition, *Cluster 2* has the largest number of patients with abnormal growth of tissue (tumor), 78%, as compared to *Cluster 1* with 60% of such patients.

We compare the performance of our method with the state-of-the-art somatic mutation-based stratification method called Network-based Stratification (NBS)[5]. NBS takes as input a

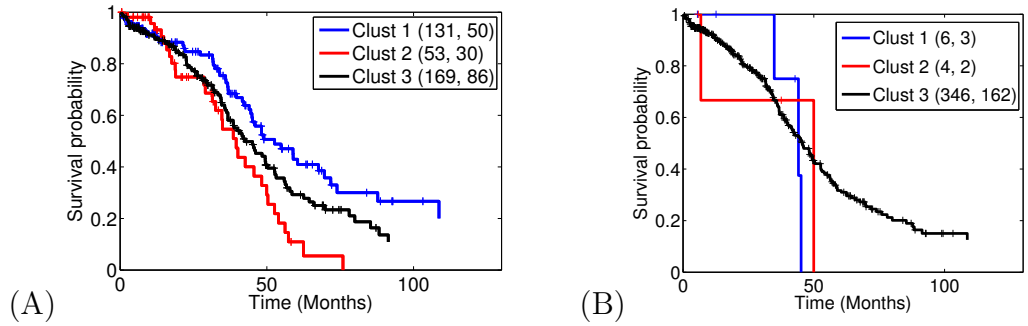

Fig. 2. Kaplan-Meier survival curves for 3 different patient groups produces by GNMTF (A) and NBS (B). The total number of patients and the number of deceased patients for each cluster are shown in the legend, the first and the second number in brackets, respectively.

patient-gene post-smoothing relation matrix and a molecular network matrix. We apply it on the same set of data described in Section 2.3, excluding drug data, which only our framework can take into account. We test NBS for different numbers of patient clusters (i.e., $k \in \{2, 3, 4, 5\}$) and compute the Kaplan-Meier survival curves[26] for the obtained patient clusters. We compare the survivability results of NBS with our method with the same number of patient clusters (Fig. 2 (A,B)). Unlike our method, which can produce clusters with significantly different survival outcomes (i.e., p-val $= 5.3 \times 10^{-3}$), NBS cannot (p-val ≥ 0.74 for all $k \in \{2, 3, 4, 5\}$). Thus, our framework is the only one able to extract personalised knowledge from somatic mutation profiles.

3.2. *Identification of driver genes*

We performed biological assessment of the $k_2 = 25$ gene clusters that we obtain from the gene consensus matrix, $\hat{\mathbf{C}}_2$. We identify 9 gene clusters that are significantly enriched in mutations and 5 gene clusters that are significantly enriched in known drivers ($p - val \leq 0.05$, see Fig. 3). Out of these clusters, cluster number 8 has the largest number of driver genes (26) and the highest enrichment in driver genes (with $p - val = 2.06 \times 10^{-4}$). To identify new driver genes, we further analyse this cluster as follows: first, based on the cluster association scores in the gene consensus matrix, we extract the mutated genes that are strongly associated with the known driver genes. In particular, we focus only on genes associated with the known driver genes with the cluster association score ≥ 0.9 (as explained below). That is, out of 20 restarts of GNMTF, we extract genes that appear 18 times in cluster 8 with other driver genes. Then, for each of these genes, we compute the average cluster score based on its associations with all driver genes. We provide the list of the top 20 genes (out of 809 predicted drivers in total) that we postulate as new driver genes of ovarian cancer progression and we sort it according to the average cluster association score, as shown in Table 1. This procedure is motivated by the observation that out of the 76 known driver genes, 67 of them are strongly related (with cluster association score ≥ 0.9) among themselves through all gene clusters.

We assess our predicted driver genes against two cancer driver gene databases, COSMIC database Cancer Gene Census[29] and IntOGen[30], as well as against a database of putative cancer driver genes, the Candidate Cancer Gene Database (CCGD)[31]. Our results show that

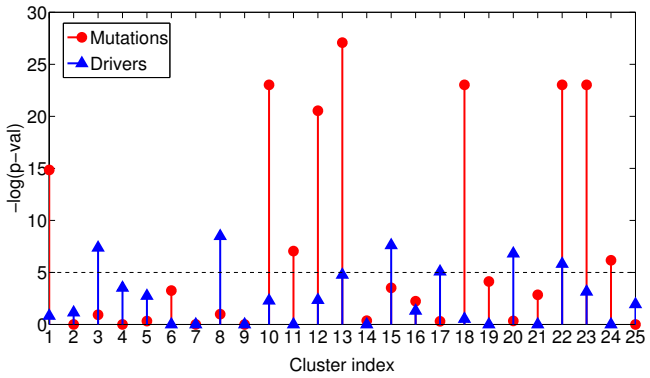

Fig. 3. Clusters enriched in mutated (red) and driver (blue) genes. For each cluster, $-log(pval)$ is plotted as a measure for enrichment (y-axis).

$\sim 40\%$ of our 809 predicted driver genes (with scores ≥ 0.9) have either been already proposed as drivers (in CCGD), or validated by experts (Census, or IntOGen). The list of our 20 top-scoring predicted cancer driver genes is presented in Table 1. Also, we investigated the literature to assess the relevance of our two top-scoring predictions that are not found in other databases and found evidence that they are biologically relevant. Our top-scoring cancer driver gene prediction is ADAM32, which is strongly clustered with driver gene BMPR2 (Table 1). The association between the two genes is biologically relevant, because both are involved with transforming growth factors (TGFs). Our prediction of ADAM32 as a cancer driver gene is also relevant, because ADAM genes are known to be responsible for cancer cell proliferation and progression[32]. The second best prediction is REG1P (from the REG family of proteins), which is strongly clustered with driver gene CLASP2. Our prediction of REG1P as a cancer driver gene is also relevant, because the REG family plays different roles in proliferation, migration, and anti-apoptosis through activating different signalling pathways; their dis-regulation is closely associated with cancer and REG proteins have been proposed as markers for prognosis of cancers[33].

3.3. *Drug-target interaction prediction*

To demonstrate the predictive power of our data fusion approach and to assess the contribution of each dataset on the drug-target interaction prediction, we perform a 5-fold cross validation for each combination of the datasets shown in Fig. 1. In all our experiments, true positives are correctly predicted DTIs, while false positives are predicted DTIs that are not present in the initial dataset.

We compute average Area Under the Receiver Operator Characteristic (ROC) and Precision-Recall (PR) curves (over 20 repetitions) to evaluate the performance of our methodology for each combination of datasets included in the integration process. The results are shown in Fig. 4. The lowest values of average AUC ROC and AUC PR are observed when only DTI dataset is used. The values increase with the inclusion of other datasets, resulting in the highest average AUC ROC when all datasets are taken into account. With all datasets taken into account by GNMTF, we use the reconstructed DTI relation matrix, $\hat{\mathbf{R}}_{23}$, to extract new drug-target interactions, as described in Sec. 2.2. We assess our prediction accuracy against two different large drug-target interaction databases, MATADOR[28] and CTD[27]. Out

Table 1. The list of the top scoring proposed driver genes (1^{st} column) and their associated known driver genes (2^{nd} column), with the association score (3^{rd} column), and the confirmation of their presence in CCGD database (4^{th} column).

New driver	Known drivers	Score	DB
ADAM32	BMPR2	1.000	–
REG1P	CLASP2	1.000	–
PCDHA2	CHD4	1.000	–
NCR1	BMPR2	1.000	–
USPL1	CLASP2	1.000	–
GDPD3	DDX5	1.000	–
LECT1	CLASP2	1.000	CCGD
IL25	CDK12, CCAR1	0.975	–
BAK1	ATRX, TFDP1, NDRG1	0.967	–
MOGAT2	ATRX, TFDP1, NDRG1	0.967	–
CHAF1A	ATRX, TFDP1, NDRG1	0.967	CCGD
PITX2	ATRX, TFDP1, NDRG1	0.967	–
SIN3B	ATRX, TFDP1, NDRG1	0.967	–
RPL30	ATRX, TFDP1, NDRG1	0.967	–
GRWD1	ATRX, TFDP1, NDRG1	0.967	–
SNAI1	ATRX, TFDP1, NDRG1	0.967	CCGD
RBMXP4	ATRX, TFDP1, NDRG1	0.967	–
CPNE7	ATRX, TFDP1, NDRG1	0.967	–
HIPK3	ATRX, TFDP1, NDRG1	0.967	CCGD
EPOR	ATRX, TFDP1, NDRG1	0.967	CCGD

Table 2. The list of predicted top scoring drug–target associations (first two columns), the association scores (third column), and the confirmation of their presence in CTD (C) or MATADOR (M) database (fifth column). All drugs are FDA-approved.

Gene	Drug	Score	Clusters	DB
KIT	ATP	0.873	1, 2, 3	–
GABRQ	Adinazolam	0.808	1	M
GABRQ	Fludiazepam	0.808	1	M
GABRQ	Cinolazepam	0.809	1	M
GABRQ	Clotiazepam	0.809	1	M
HTR2A	Dopamine	0.809	1, 3	C, M
GRIN3A	Pethidine	0.801	1, 2	–
CACNA2D1	Verapamil	0.761	1, 3	M
PDGFRB	ATP	0.724	1, 2	–
KDR	ATP	0.724	1, 3	C
HTR1A	Mirtazapine	0.720	1, 2	C ,M
GABRA6	Adinazolam	0.688	1	M
GABRA6	Fludiazepam	0.688	1	M
GABRA6	Cinolazepam	0.688	1	M
GABRA6	Clotiazepam	0.688	1	M
GABRA4	Adinazolam	0.687	1, 3	M
GABRA4	Fludiazepam	0.687	1, 3	M
GABRA4	Cinolazepam	0.687	1, 3	M
GABRA4	Clotiazepam	0.687	1, 3	M
CACNA1D	Magnesium Sulfate	0.676	1, 2, 3	M

of our 225,947 predicted DTIs, 37% have already been confirmed in MATADOR, or CTD. The list of our 20 top scoring predicted DTIs is shown in Table 2, out of which 17 are confirmed in CTD, or MATADOR database. Second, we investigated the literature to assess the relevance of our two top-scoring predicted DTIs that are not found in other databases and found evidences that they are biologically relevant. The top scoring target gene KIT (C-Kit) is particularly relevant. It is a receptor tyrosine kinase (e.g., it catalyses ATP/ADP reactions). It has been shown that unregulated activity of this gene leads to occurrence of tumors and thus, it has been proposed as a potential drug target in cancer[34]. Interestingly, we predict the drug candidate for targeting this gene to be Adenosine triphosphate (or ATP), for which a precise role in cancer is still under investigation. Increasing ATP intake is known to improve cancer patient conditions[35]. The reason could be that ATP is linked to cancer cell metabolism and either activates cell death mediated by restoration of normal mitochondrial function, or alterates the cytosolic ATP/ADP ratio, which is postulated to deactivate glycolysis (Warburg effect) in a cancer cell[36]. Another drug-target in Table 2 whose predicted drug is not present in CTD and MATADOR databases is GRIN3A. GRIN3A (NMDAR-l) is a sub-unit of NMDA receptor (a glutamate-regulated ion channel). NMDA receptor has been proposed as a target for cancer chemotherapy[37]. It has been proposed that glutamate antagonist molecules should be used as potential drug targets[37]. Interestingly, our predicted drug, Pethidine (also known as Meperidine), is a glutamate antagonist that is known to bind NMDA receptors[38], which provides evidence that our prediction of Pethidine as a drug for targeting GRIN3A is biologically relevant. However, evidence that Pethidine can bind to GRIN3A in particular has not yet been established. Furthermore, based on the mutated genes of particular patients, we propose these newly discovered drugs (see column four in Table 2) for treatment of the three patient groups described in Sec. 3.1.

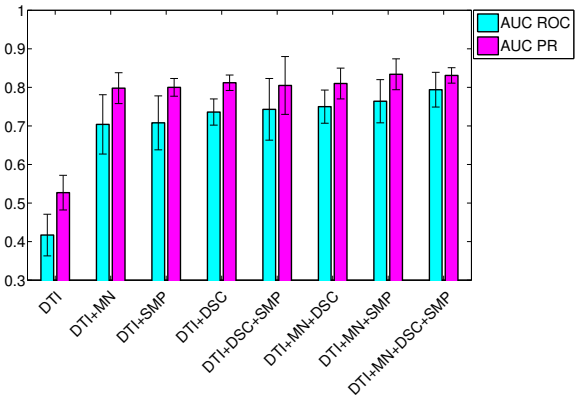

Fig. 4. Area under the ROC and PR curves for GNMTF runs done on the combination of particular datasets listed on x-axis. The results are sorted increasingly according to the AUC ROC values. See Fig. 1 for the abbreviated names of the datasets.

4. Conclusions

In this paper, we propose a data fusion framework that can effectively integrate somatic mutation data along with molecular networks and drug chemical data. It is based on GNMTF method for co-clustering heterogeneous data and it can be even further extended to incorporate any number and type of data. One important advantage of our framework is that when applied to a specific cancer, it can simultaneously perform three different tasks: patient stratification into clinically different groups, novel driver gene identification and drug-repurposing predictions for treating cancer.

We apply the GNMTF-based data fusion framework to ovarian cancer patients and identify three substantially different groups of patients with different survival outcomes. In addition, from the obtained gene clusters, we identify a list of genes that we postulate as potential drivers of ovarian cancer progression due to their strong cluster associations to known ovarian cancer driver genes. We perform biomedical literature curation for the top scoring predictions, ADAM32 and REG1P, and show that they are related to cancer cell proliferation and tumor progression, while 40% of other predictions we validate in other databases. Moreover, our framework is capable of predicting new drugs that could be used for targeting mutated genes and thus, for treatment of identified groups of ovarian cancer patients. We provide a list of predicted drug-target interactions, a good number of which is matching those reported in other databases. Other, non-validated predictions for driver genes and drug-target interactions could be true, awaiting experimental validation.

Our analysis also suggests that somatic mutation data is a valuable complement to other molecular data, whose integration with those data could lead to an improvement in the performance of data fusion methods. Our approach has a potential to enable the derivation of new hypotheses, improve drug selection and lead to improvement in patient genomics-tailored therapeutics for cancer.

Acknowledgement

This work was supported by the European Research Council (ERC) Starting Independent Researcher Grant 278212, the National Science Foundation (NSF) Cyber-Enabled Discovery and Innovation (CDI) OIA-1028394, the ARRS project J1-5454, and the Serbian Ministry of

Education and Science Project III44006.

References

1. B. Vogelstein *et al.*, *Science* **339**, 1546 (2013).
2. C. Kandoth, M. D. *et al.*, *Nature* **502**, 333 (2013).
3. C. Rubio-Perez, *et al. Cancer Cell* **27**, 382 (2015).
4. Cancer Genome Atlas Research Network, *Nature* **474**, 609 (2011).
5. M. Hofree *et al.*, *Nature Methods* **10**, 1108 (2013).
6. Cancer Genome Atlas Research Network, *Nature* **487**, 330 (2012).
7. Y. Chen *et al.*, *Scientific Reports* **3** (2013).
8. Y. Yamanishi *et al.*, *Bioinformatics* **26**, i246 (2010).
9. C. Ding *et al.*, in *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, (2006).
10. C. Ding *et al.* in *Proceedings of SIAM Data Mining Conference*, (2005).
11. F. Wang, T. Li and C. Zhang, in *SIAM Conference on Data Mining (SDM)*, (2008).
12. M. Žitnik *et al. Scientific Reports* **3** (2013).
13. M. Žitnik and B. Župan, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **37**, 41 (2015).
14. T. Hwang *et al.*, *Nucleic Acids Research* **40**, e146 (2012).
15. H. Wang *et al.*, *Journal of Computational Biology* **20**, 344 (2013).
16. V. Gligorijević, V. Janjić and N. Pržulj, *Bioinformatics* **30**, i594 (2014).
17. A. Chatr-aryamontri *et al.*, *Nucleic Acids Research* **43**, D470 (2015).
18. M. Kanehisa and S. Goto, *Nucleic Acids Research* **28**, 27 (2000).
19. D. S. Wishart, C. Knox, A. C. Guo *et al.*, *Nucleic Acids Research* **36**, D901 (2008).
20. F. Shang, L. Jiao and F. Wang, *Pattern Recognition* **45**, 2237 (2012).
21. R. Albright, J. Cox, D. Duling, A. Langville and C. Meyer, *North Carolina State University, Tech. Rep* **81706** (2006).
22. J.-P. Brunet *et al.*, *Proceedings of the National Academy of Sciences* **101**, 4164 (2004).
23. H. Kim and H. Park, *Bioinformatics* **23**, 1495 (2007).
24. O. Vanunu *et al.*, *PLoS Computational Biology* **6**, e1000641 (01 2010).
25. N. Nikolova and J. Jaworska, *QSAR & Combinatorial Science* **22**, 1006 (2003).
26. V. Bewick, L. Cheek and J. Ball, *Critical Care* **8**, 389 (2004).
27. A. P. Davis *et al.*, *Nucleic Acids Research* **43**, D914 (2015).
28. S. Günther *et al.*, *Nucleic Acids Research* **36**, D919 (2008).
29. P. A. Futreal *et al.*, *Nature Reviews Cancer* **4**, 177 (2004).
30. G. Gundem *et al.*, *Nature Methods* **7**, 92 (2010).
31. K. L. Abbott *et al. Nucleic Acids Research* **43**, D844 (2015).
32. S. Mochizuki and Y. Okada, *Cancer Science* **98**, 621 (2007).
33. J. Zhao, J. Wang, H. Wang and M. Lai, *Adv. Clin. Chem* **61**, 153 (2013).
34. J. Lennartsson and L. Ronnstrand, *Current Cancer Drug Targets* **6**, 65 (2006).
35. A. Jatoi and C. L. Loprinzi, *Journal of Clinical Oncology* **20**, 362 (2002).
36. E. N. Maldonado and J. J. Lemasters, *Mitochondrion* **19**, 78 (2014).
37. S. I. Deutsch, A. H. Tang, J. A. Burket and A. D. Benson, *Biomedicine & Pharmacotherapy* **68**, 493 (2014).
38. T. Yamakura, K. Sakimura and K. Shimoji, *Anesthesia & Analgesia* **90**, 928 (2000).