JAMIA Open, 1(1), 2018, 75–86
doi: 10.1093/jamiaopen/ooy008
Advance Access Publication Date: 14 May 2018
Research and Applications

Research and Applications

Patient similarity by joint matrix trifactorization to identify subgroups in acute myeloid leukemia

F. Vitali,[1,2,3,‡] S. Marini,[4,‡] D. Pala,[5] A. Demartini,[5,6] S. Montoli,[5,6] A. Zambelli,[7] and R. Bellazzi[5,6,8]

[1]Center for Biomedical Informatics and Biostatistics, The University of Arizona, Tucson, Arizona, USA, [2]BIO5 Institute, The University of Arizona, Tucson, Arizona, USA, [3]Department of Medicine, The University of Arizona, Tucson, AZ, USA, [4]Department of Computational Biology and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA, [5]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, PV, Italy, [6]Centre for Health Technologies, University of Pavia, PV, Italy, [7]Oncology Unit, ASST Papa Giovanni XXIII, Bergamo, BG, Italy and [8]IRCCS Istituti Clinici Scientifici Maugeri, Pavia, PV, Italy

Corresponding Author: Dr. Riccardo Bellazzi, University of Pavia, Department of Electrical, Computer and Biomedical Engineering, 27100, Pavia, PV, Italy (riccardo.bellazzi@unipv.it)
[‡]These authors contributed equally to the work.

Received 20 December 2017; Revised 7 March 2018; Accepted 20 March 2018

ABSTRACT

Objective: Computing patients' similarity is of great interest in precision oncology since it supports clustering and subgroup identification, eventually leading to tailored therapies. The availability of large amounts of biomedical data, characterized by large feature sets and sparse content, motivates the development of new methods to compute patient similarities able to fuse heterogeneous data sources with the available knowledge.
Materials and Methods: In this work, we developed a data integration approach based on matrix trifactorization to compute patient similarities by integrating several sources of data and knowledge. We assess the accuracy of the proposed method: (1) on several synthetic data sets which similarity structures are affected by increasing levels of noise and data sparsity, and (2) on a real data set coming from an acute myeloid leukemia (AML) study. The results obtained are finally compared with the ones of traditional similarity calculation methods.
Results: In the analysis of the synthetic data set, where the ground truth is known, we measured the capability of reconstructing the correct clusters, while in the AML study we evaluated the Kaplan-Meier curves obtained with the different clusters and measured their statistical difference by means of the log-rank test. In presence of noise and sparse data, our data integration method outperform other techniques, both in the synthetic and in the AML data.
Discussion: In case of multiple heterogeneous data sources, a matrix trifactorization technique can successfully fuse all the information in a joint model. We demonstrated how this approach can be efficiently applied to discover meaningful patient similarities and therefore may be considered a reliable data driven strategy for the definition of new research hypothesis for precision oncology.
Conclusion: The better performance of the proposed approach presents an advantage over previous methods to provide accurate patient similarities supporting precision medicine.

Key words: data integration, matrix trifactorization, acute myeloid leukemia (AML), patient similarity, precision medicine

BACKGROUND AND SIGNIFICANCE

The concept of precision medicine is based on the assumption that a careful identification of patients' subgroups is able to properly take into account individual variability, which may play a major role in any prevention and treatment strategies.[1] This concept is not new: blood typing, for instance, has been used to correctly allocate blood transfusions for more than a century.[2]

The oncology-field seems to be a clear choice for taking advantage of precision medicine. Cancers are common diseases highly impacting on the population because of their lethality, severe symptoms, and toxicity associated with the oncological treatment. Moreover, each cancer has its own genomic signature, along with some common features shared by multiple cancer types.[3] *Patient similarity* is an emerging approach in precision oncology and medicine, identifying patients with similar profiles and derive insights to investigate diseases and potential treatments. In precision oncology, patient similarity is traditionally measured through preidentified signatures (eg Oncotype DX$^{\mathrm{TM}}$,[4] PAM50,[5] and other clinically available classifiers), or patient-specific biomarkers.[6] However, these preidentified onco-signatures rely only on a relatively small number of molecular features, and the trials launched to test the real impact of the precision oncology in daily clinical practice have so far yield little results,[7–9] being limited only to a tiny proportion of the entire population enrolled in the trials.[9–11] Therefore, conventional studies designed on the basis of the classical 4-phases of drug-development are probably neither effective nor fit for precision oncology.[12]

The availability of large heterogeneous biomedical data naturally opens ways to develop computational methods leveraging on the whole multidimensional patient data framework to search for patient similarity. These data include, among others, clinical (ie coded data, text, images, signals), -omics (from genome to metabolome), and exposome data. The capability of computing patient similarity in presence of such large features becomes therefore a crucial component to enable large-scale precision medicine implementation.[13] In literature, patient similarity seems highly dependent on the specific problems considered, and there is no consensus about the best metrics or the best algorithms to calculate it in presence of heterogeneous and sparse data.[14] To face these problems and consider patient similarity from a multidimensional perspective, in recent years a number of methods to determine patient similarity has been developed.[15–22]

In this article, we propose a novel method to compute patient similarity for precision oncology by an unsupervised discovery of patient subgroups. This method is based on a strategy that integrates data and knowledge in a sound and formal way. In particular, we exploited a modified version of a non-negative matrix trifactorization algorithm recently developed[23] and applied also to biomedical problems by Gligorijevic et al,[21] Zang et al,[24] Utro et al, and Zitnik et al.[21,25–30] Factorization techniques are efficient tools for data fusion of large sparse data sets (like the ones available in the clinical setting). These approaches adopt a useful dimension reduction by directly compressing the starting data into a lower number of features (ie vector components). The decomposition methods play a central role in the analysis of the latent structure hidden in the data that may unveil unknown interactions of the initial data, that is patient similarities. In recent years, several dimension reduction techniques have been proposed to tackle biological problems. Examples rely on collective matrix factorizations,[31,32] tensor decomposition,[33] Bayesian multitensor factorization,[34] and group factor analysis (GFA).[35,36]

Our approach takes into account the structural relation of several highly heterogeneous data, such as clinical and genomic data, and available knowledge from several public repositories. It implements both metafeatures extraction and distance measures to reveal hidden patient similarities. In the majority of state-of-art dimension reduction methods, there is no constraint on the sign of the metafeature elements, thus admitting negative components or subtractive combinations into the representation.[37] Our method, on the contrary, is based on non-negative trifactorization. The incorporation of non-negative constraints has been shown to enhance the interpretability of the data integrated.[37] We tested the performance of the proposed algorithm on different synthetic data sets, affected by increasing levels of noise and data sparsity. We further validated the method by fusing a real data set coming from an AML study with several external knowledge sources. By comparing it with state-of-art techniques, we show our method outperforms other approaches in both simulated and real data. Identified patients' subgroups are validated as significantly different by survival curves.

MATERIALS AND METHODS

An overview of the trifactorization algorithm

The structure of data sources and knowledge bases is typically organized into relational matrices associating various objects/concepts, such as patients, clinical data, genes, diseases, and so on. The non-negative trifactorization algorithm naturally exploits these data structures to perform data fusion by first representing them in a matrix form and subsequently organizing them in a unique big block matrix. The algorithm aims at identifying low-rank non-negative matrices whose product can provide a good approximation of the original non-negative matrix. The result is a new matrix containing predictions and novel knowledge about the associations represented. This algorithm can be considered as a knowledge-based method that allows dealing with sparsity by interpolating missing data through a prediction derived by explicitly modeling the correlation and the dependency between attributes.

A Matlab implementation of our algorithm is available at https://gitlab.com/smarini/MaDDA.

The algorithm is described step by step as follows.

Let us consider r different types of concepts, say, patients, genes, miRNAs, …, which we call *objects* $o_1, o_2, \dots, o_r$ and let's suppose that we have a set of data sources that relate pairs of objects (o_i, o_j) for some i and j: for example we can have the objects "gene" and "disease" and the repository "DisGeNeT"[38] that relates them. If the number of objects of type o_i are n_i and the number of objects of type o_j are n_j the data source when $i \neq j$ can be represented as a sparse matrix $R_{ij} \in \mathbb{R}^{n_i \times n_j}$, called *relation matrix* (Figure 1). For instance, the relation matrix may contain information of the relationships between genes (eg BRCA1) and diseases (eg breast cancer). If we also have observations about the relationships of the objects of the same type, such as genes coexpression, we might represent them with a matrix $\Theta_i \in \mathbb{R}^{n_i \times n_i}$, called *constraint matrix* (Figure 1).

Considering the entire set of R_{ij} relation matrices given by all the data sources of interest, we can represent them as a block matrix R, which may miss elements (eg not all the genes in the genome can be related to a given disease):

$$R = \begin{pmatrix} * & R_{12} & \dots & R_{1r} \\ R_{21} & * & * & R_{2r} \\ \vdots & \vdots & \ddots & * \\ R_{r1} & R_{r2} & \dots & * \end{pmatrix} \qquad (1)$$

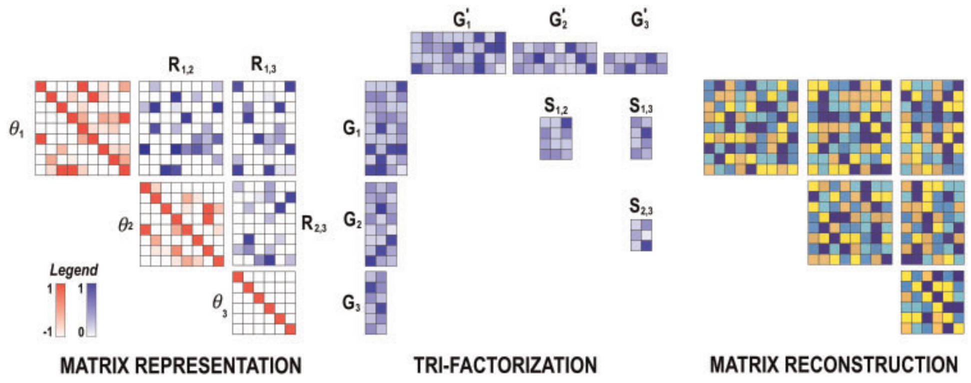


Figure 1. An example of the trifactorization algorithm constructed by considering 3 data sources. All the data sources are represented as relation matrices. $R_{i,\ j}$ matrices are used to describe associations between objects of the different type (eg gene-disease), their values range between [0, 1], where 1 indicates strong interaction and 0 association absence or lack. Θ_i matrices represent relations between objects of the same type (eg gene-gene) and Θ_i elements vary between [−1, 1], where −1 represents a strong association and 1 a lack of association. R matrix is then trifactorized by running and optimization algorithm (see Eq. 5) into a set of lower-rank factors, eg G_i and $S_{i,\ j}$. Finally, the whole matrix is reconstructed by multiplying matrices G_i and $S_{i,\ j}$, thus revealing new associations.

The values of the matrix R express the strength of the relationships between objects, and they correspond to numbers between 0 and 1, with 0 meaning no known relationships.

On the other hand, the constraint matrices Θ_i can be expressed as a diagonal block set, Θ_i^t, where $t = 1, 2,\ldots, i$ denotes the possible multiplicity of relationships of the same type, which can be derived by i different knowledge sources (corresponding to different Θ_i matrices of the same object). For example, coexpression may be measured through different types of experiments.

$$\Theta^{(t)} = \mathrm{Diag}\left(\Theta_1^{(t)}, \Theta_2^{(t)}, \ldots, \Theta_r^{(t)}\right). \tag{2}$$

Differently from R_{ij} matrices, Θ_i values vary between −1 and 1, expressing the dissimilarity between elements of the same object types, so that −1 means full similarity while 1 is full dissimilarity.

Once the data are represented into matrices, the trifactorization algorithm jointly factorizes the matrices R_{ij} using the matrices Θ_i as constraints. First of all, a set of design parameters $k_i \ll n_i$ is defined for each object. These parameters, also called *ranks*, define the dimension of the latent factors for the ith object type with the objective of revealing hidden structure in the data. This is a crucial step in the algorithm, since wrongly assigned ranks may lead to overfitting (if too big), or may not be able to capture all the information (if too small). There is no general consensus about how to select these values and different approaches can be applied.[21,26,30] In this work, we opted for an empirical approach (see Selection of initial parameters section).

After rank selection, each block of the matrix R is factorized in 2 lower-rank block matrices, G and S (Figure 1), as follows:

$$G = \mathrm{Diag}(G_1^{n_1 \times k_1}, G_2^{n_2 \times k_2}, \ldots, G_r^{n_r \times k_r}) \tag{3}$$

$$S = \begin{pmatrix} * & S_{12}^{k_1 \times k_2} & \ldots & S_{1r}^{k_1 \times k_r} \\ S_{21}^{k_2 \times k_1} & * & * & S_{2r}^{k_2 \times k_r} \\ \vdots & \vdots & \ddots & * \\ S_{r1}^{k_r \times k_2} & S_{r2}^{k_r \times k_2} & \ldots & * \end{pmatrix} \tag{4}$$

Thanks to the joint factorization, information is spread out from and to the different relation matrices, so that the missing elements are predicted on the basis of the multiplication of elements of much lower rank.

The matrices S and G are reconstructed by minimizing the following objective function:

$$\min_{G \geq 0} J(G; S) = \sum_{R_{ij} \in R} \|R_{ij} - G_i S_{ij} G_j^t\|^2 + \sum_{t=1}^{\max_i t_i} tr(G^t \Theta^{(t)} G), \tag{5}$$

where $|| \cdot ||$ is the Frobenius norm and tr() the trace of the matrix. The procedure adopted to solve Eq. 5 starts with a random initialization of the G, next, S matrices are iteratively updated until convergence (proof of convergence and details in references 23 and 30). Details on the adopted procedure to solve the optimization problem provided in Supplementary Material Method S1.

The algorithm described above has been adapted to calculate the similarity between the same type of objects: in this article, we are interested in the object "patients." With a closer look to the approximation $R_{ij} \approx G_i S_{ij} G_j^t$, we can notice that matrix G_i is shared by all blocks that are related to the object type o_i (in our case patients), while S_{ij} is specific to the relations between the objects o_i and o_j. Since G_i is an $n_i \times k_i$ matrix, the rows correspond to the elements of the ith object type (in our case the different patients), while the columns represent k_i groups. Therefore, each element can be interpreted as the degree of membership of each patient (row) to each group (column). Therefore, we can assign an element (ie a patient) to the group (ie to a cluster) with the largest value, that is the column with the maximum value for the corresponding row.

Since the optimization strategy strongly depends on the initialization (ie the selection of the dimension of k_i parameters), we averaged the results over 10 applications to obtain a final consensus matrix ($\hat{C}$), which is calculated as the element-wise averaged sum of the connectivity matrices. In our case, for example, a consensus matrix element showing a value of 0.5 means that 5 times out of 10, in the connectivity matrix, the 2 patients corresponding to the row and column indexes of the element ended up grouped together.

Patient and external knowledge-based data sets

In Table 1 are reported the data sources used in this work. In detail, we propose to integrate patient data (extracted from TCGA,[39]) and several publicly available knowledge bases to extract meaningful patient similarities.

Table 1. Patient data and external knowledge sources

Data type	Retrieved data	Resource[b]
Patient data on acute myeloid leukemia		
1) Clinical data (200 patients)	Gender, age, vital status: dead or alive, days to death (if dead), days to birth, days to last follow-up, date of the diagnosis[a]	TCGA[c,39]
2) Somatic mutations (195 patients)	1620 mutations associated with 428 genes	
3) Gene expression profiles (197 patients)	22578 genes (8897 after filter application)	
External knowledge data sources		
4) Gene-gene interactions	Starting from the 186 genes involved in AML (extracted from MalaCard) and the 428 genes associated with the mutations, we extracted 37 811 first-degree gene-gene interactions between 8897 unique genes.	BioGRID[d,40]
5) Gene-pathway associations	3202 associations between the 8897 genes and 383 KEGG pathways	KEGG[e,41]
6) Disease-disease relationships	35 201 associations between 6402 unique diseases.	DO[f,42]
7) Disease-gene associations	1925 associations between 6402 diseases and 278 genes.	DisGeNET v4.0[38]
8) Disease-pathway relations	605 associations between the 6402 diseases and the 383 pathways.	KEGG[e,41]

[a] We listed the clinical variables used in this work.

[b] Data have been accessed in November 2016.

[c] The Cancer Genome Atlas. Data have been retrieved by using the cBioPortal for Cancer Genomic,[43] last updated 05/31/16.

[d] Biological General Repository for Interaction Datasets, Release 3.4.142.

[e] Kyoto Encyclopedia of Genes and Genomes, Release 80.0.

[f] Disease Ontology, Release 2016-01-07.

The AML TCGA cohort was used to (1) generate the data set for a simulated study and (2) to validate the proposed approach on a real data set. Gene expression data were normalized by using robust multichip average (RMA),[44] normalization method. Mutation data were analyzed using the software PaPi.[45] PaPi is a machine-learning approach to classify and score human coding variants by estimating the probability to damage their protein-related function. Each of the 1620 mutations,[45] gets a score (probability) between 0 (ie tolerated mutation) and 1 (ie damaging mutation). A list of 186 genes involved in AML were selected from MalaCards.[46] This list was subsequently integrated with the 428 genes associated with the mutation data (Table 1). We finally retained the gene expression and mutation data for that list of ~8900 genes and for the genes extracted as first-degree interactor in BioGRID.[40]

Matrix definition

We represented all the data sources (Table 1) in the form of relation matrices, each one formalizing the known associations between pairs of objects. In detail, the considered objects are: o_1 clinical data; o_2 mutations; o_3 genes; o_4 pathways; o_5 diseases; and o_6 patients. According to the integration algorithm and in order to make the matrices comparable, associations between objects of different type (R_{ij} matrices) were rescaled in the interval [0, 1]. On the other hand, association between objects of the same type (Θ_i matrices) were rescaled in the interval [−1, 1].

Details on data processing are provided in Table 2. Matrix data are available at https://gitlab.com/smarini/MaDDA.

Selection of initial parameters

A crucial step in the factorization algorithm is the selection of the input ranks k_i. All the ranks, except the one related to the patient object (k_6), were computed resorting to an empirical rule proposed in reference 29 (Eq. 10).

$$\underset{i \neq 6}{k_i = \frac{1}{200}\frac{\left(nnz_i^{\text{rows}} + nnz_i^{\text{cols}}\right)}{2}}, \tag{10}$$

where nnz are the nonzero elements of the object i counted on the rows (nnz_i^{rows}), and on column (nnz_i^{cols}), respectively.

On the other hand, the patient rank k_6 corresponding to the number of expected clusters is computed following the approach presented in reference 21. We applied a grid search to select k_6, since Eq. 10 would provide a very high rank (ie high number of patient clusters) due to the low sparsity of the gene/patient relationships. Different values of k_6 from a predefined interval are used as inputs to the integration algorithm, and the results are compared in terms of their dispersion coefficient, the larger the better (Eq. 11):

$$\rho\left(\hat{C}^{(k_6)}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} 4*[\hat{C}_{ij}^{(k_6)} - \frac{1}{2}]^2 \quad l_{k_6} = [3,\ 5,\ 10,\ 20] \tag{11}$$

where k_6 is a patient rank value from the list l, $\hat{C}^{(k_6)}$ is the consensus matrix (see An overview of the trifactorization algorithm section) computed by using k_6. The rank k_6 obtaining the higher $\rho(\hat{C}^{(k_6)})$ value was used as rank input for the proposed approach. This procedure has been used for both the synthetic and the real data sets.

Simulated data set construction

To evaluate the performance of the proposed approach, we generated different synthetic data sets with the same size of the real one (Table 1).

By varying 2 simulation parameters listed in Supplementary Material Table S1, we created 25 data sets, differing for amount of added noise, missing data, degrees of patient similarity, and data sparsity. For each scenario, we simulated 200 virtual patients grouped into 5 clusters with known similarity structures. The simulation process is as follows:

1. Patient clinical data, gene expression levels, and the PaPi score mutation data are used to construct a patient similarity matrix by computing the Euclidean distance (ED) between patients.

Table 2. Matrix R_{ij} and Θ_i construction

Matrix	Relation	Matrix value definition
$R_{16} = R_{61}$	Clinical data-patient**	We used 4 clinical variables (ie rows) for each patient (ie column), that is *Female*, *Male*, *Age*, and *Survival*. The gender variable was used to create 2 rows, that is *Female* and *Male*, for each patient, whose value was set to 0 or 1 according to the gender. The age field was used to define the *Age* row whose values are given by: $\text{age}\ (i) = \frac{a_i - a_{\min}}{a_{\max} - a_{\min}}$ (6) where a_i is the *i*th patient age at the time of the death or at the last follow-up, while $a_{\min}$ and $a_{\max}$ are the cohort minimum and maximum ages, respectively. Finally, a *Survival* row for each patient was obtained by dividing the patients into alive and deceased individuals. Since the majority of the patients died within 1 year from the diagnosis date, we considered the survival *S* at one year computed as: $S(i) = \begin{cases} 1 & \text{if } d_i > 365 \text{ days} \\ \dfrac{d_i - d_{\min}}{d_{\max} - d_{\min}} & \text{if } d_i < 365 \text{ days and vital status} = \text{deceased} \\ 0 & \text{if } d_i < 365 \text{ days and vital status} = \text{alive} \end{cases}$ (7) where d_i is the number of days between the diagnosis and death rate or the follow-up date of the patient i. $d_{\min}$ and $d_{\max}$ are the global minimum and the maximum number of days between the diagnosis and death date, respectively. 0 means that it is unknown if the patient is alive or deceased 1 year after the diagnosis since the last follow-up date occurred before that time. This allowed to obtain values ranging between [0, 1].
$R_{23} = R_{32}$	Mutation-Gene	Mutations mapped to their respective genes.
$R_{26} = R_{62}$	Mutation-patient	PaPi[45] scores evaluating harmfulness of each mutation, per patient.
$R_{34} = R_{43}$	Gene-disease	We used DisGeNET data associating genes and diseases. Since DisGeNET score provided is already in the interval [0, 1], no further processing was required.
$R_{35} = R_{35}$	Gene-pathway*	Genes mapped to KEGG pathways. Presence/absence of a gene in pathway determines its binary 1/0 value.
R_{36}	Gene-patient	Gene-patient values correspond to the sum (mutational burden) of the PaPi scores of all the mutations associated to a specific gene. The obtained values were then rescaled between [0, 1].
$R_{45} = R_{54}'$	Disease-pathway*	Matrix values correspond to 0 or 1 according to the information extracted from KEGG about all the diseases altering each KEGG pathway.
$R_{46} = R_{64}$	Disease-patient	Rows of this matrix represent the diseases and column the 200 patients. The values of the row corresponding to acute myeloid leukemia (DOID: 9119) were set to 1 indicating the association.
R_{63}	Patient-gene	Gene expression data from TCGA were used as matrix values according with the formula: $\text{Expression}(i,j) = \frac{e_{i,j} - e_{\min}}{e_{\max} - e_{\min}}$ (8) where e_i is the patient *i* expression of the gene *j*, while $e_{\min}$ and $e_{\max}$ are the global minimum and the maximum values of the gene *j*.
Θ_1	Clinical data-clinical data	The rows and the columns of this matrix are *Age, Female, Male*, and *Survival*. The diagonal values are −1 (ie fully associated). *Male* and *female* association is 1 (ie not associated).
Θ_1	Mutation-mutation	No assumption was made on the mutation similarity, beside considering each mutation similar to itself.
Θ_3	Gene-gene	Gene-gene interactions from BioGRID. The raw data needed a preprocessing step, since for each gene pair, the related associations may appear multiple times (corresponding to different kind of interaction, eg direct physical binding, genetic interaction[40]). For this reason, denoting with x the number of times a certain pair appears, its score was determined by: $f(x) = -\frac{1}{2}\left(1 + \frac{\ln(x)}{\ln(x_{\max})}\right)$ (9)
Θ_4	Diseasedisease	The similarity between 2 diseases is set to -0.8^n where *n* is the length of the shortest path between corresponding terms in the DO (ie the minimum number of steps to reach one disease from the other one).
Θ_5	Pathway-pathway	KEGG relations were used to measure binary pathway similarity. Also, each pathway is considered similar to itself.
Θ_6	Patient-patient	No assumption was made on the patient similarity, besides considering each patient similar to itself.

[a] Data have been automatically extracted by using Python Bioservice 1.4.8.[47]

[b] In case of missing data, the value of the element was set to 0 (ie unknown).

Next, we selected the 5 less similar patients according with the mean ED values.

2. For each patient P of the 5 patients identified in (i), 39 'virtual patients' are generated by adding Gaussian noise with mean of zero and variance equal to one half of the population variance to all the *P*'s clinical data (excluding the gender variable). The gender is assigned according with a probability = .9 to be the same as P. Finally, gene expression levels and PaPi mutation scores of a new simulated patient *j* are obtained according with Eq. 12.

$$D_i^j = D_i^0 \pm \mathrm{CV}\ m_i, \qquad 1 < j < 39 \tag{12}$$

where D_i^0 is the patient j's value of the object i (gene or mutation) in the reference data set, the value m_i is the mean value of an element in the population of the object i (eg mean expression value of a gene), and CV is the coefficient of variation.[48]

3. A percentage d_i sparsity is added to all the simulated data set in order to test the robustness of the approach to sparse data.

With this procedure, we generated the 25 simulated data sets by varying the CV parameter and the data sparseness as reported in Supplementary Material Table S1. The obtained data were then integrated with the external knowledge (Table 1) in order to apply the proposed algorithm.

Result evaluation of synthetic data

We evaluated the performance of the trifactorization algorithm on the 25 simulated data sets by measuring the mean absolute error (MAE),[49] defined as:

$$\mathrm{MAE} = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{1}{N} \left| \widehat{C_{ij}} - C_{ij}^{\text{expected}} \right|, \tag{13}$$

where n and m are the number of matrix rows and columns, N is the total number of matrix elements, $\hat{C}_{ij}$ is the estimated consensus matrix in the position ij, $C_{ij}^{\text{expected}}$is the ideal consensus matrix (ie each simulated patient is clustered with its corresponding real patient, for a total of 5 clusters of 40 patients).

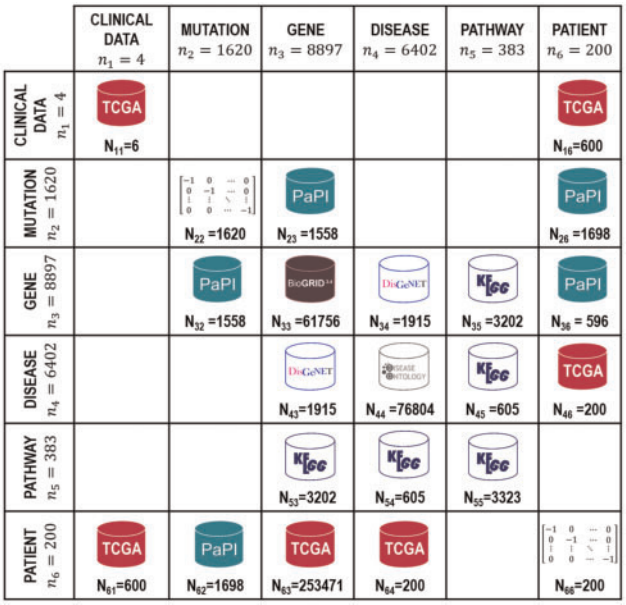
Result comparison with other techniques

In addition, using the simulated data sets, we assessed and compared the proposed approach with 2 widely used methods, that is principal component analysis (PCA),[50] and the ED measure. The results were also compared to 2 more advances techniques to integrate heterogeneous data: (1) a deep learning method based on restricted Boltzmann machines,[51] that is multimodal deep belief networks (MDBN),[20,52] and (2) a GFA,[36] based on factor analysis. Unlike the matrix trifactorization algorithm, these methods are not designed to include external knowledge sources associating entities not related to the prediction target (ie gene-gene interactions do not involve patients). These methods were therefore applied considering only the patient-related data ($R_{1,\ 6}$; $R_{2,\ 6}$; $R_{3,\ 6}$; $R_{6,\ 3}$; Table 2). Details on the adopted procedures used to apply PCA, ED, MDBN, and GFA methods are provided in Supplementary Material Method S1.

In order to evaluate their performances, MAE was computed by replacing in Eq. 12 the $\hat{C}_{ij}$ matrix with the similarity matrix Sim_{ij} obtained by applying the PCA, EA, MDBN, or GFA method. In this case, Sim_{ij} is built by setting its elements $\mathrm{Sim}_{ij} = 1$ if the patient i and the patient j belong to the same cluster, and $\mathrm{Sim}_{ij} = 0$, otherwise.

Validation case study of acute myeloid leukemia

We investigated the biological relevance of the patient similarities uncovered by the proposed methodology focusing on AML. AML is a myeloid neoplasm related to an uncontrolled proliferation of white blood cells. It is the most common leukemia in adult patients. AML is curable from 35 to 40% of patients younger than 60 years of age, while for older patients the percentage decreases from 5 to 10%.[53,54] Moreover, AML is an heterogeneous disease and, over the past 15 years, its molecular heterogeneity has become apparent,[54] thanks to -omics technologies. The analysis of public data on AML combined with external knowledge sources may reveal novel insights into AML patient profiles to discriminate between good or bad responders and suggest tailored therapies. In order to test our approach, we collected patient data from The Cancer Genome Atlas (TCGA),[39] and integrated them with external knowledge sources (Table 1). Collected data were then transformed into a matrix format according with Table 2.



	CLINICAL DATA $n_1 = 4$	MUTATION $n_2 = 1620$	GENE $n_3 = 8897$	DISEASE $n_4 = 6402$	PATHWAY $n_5 = 383$	PATIENT $n_6 = 200$
CLINICAL DATA $n_1 = 4$	TCGA N_{11}=6					TCGA N_{16}=600
MUTATION $n_2 = 1620$		$\begin{bmatrix} -1 & 0 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 \end{bmatrix}$ N_{22} =1620	PaPI N_{23} =1558			PaPI N_{26} =1698
GENE $n_3 = 8897$		PaPI N_{32} =1558	BioGRID N_{33} =61756	DisGeNET N_{34} =1915	KEGG N_{35} =3202	PaPI N_{36} = 596
DISEASE $n_4 = 6402$			DisGeNET N_{43}=1915	DISEASE ONTOLOGY N_{44} =76804	KEGG N_{45} =605	TCGA N_{46} =200
PATHWAY $n_5 = 383$			KEGG N_{53}=3202	KEGG N_{54}=605	KEGG N_{55}=3323	
PATIENT $n_6 = 200$	TCGA N_{61}=600	PaPI N_{62}=1698	TCGA N_{63}=253471	TCGA N_{64}=200		$\begin{bmatrix} -1 & 0 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 \end{bmatrix}$ N_{66}=200

Figure 2. Data sources and matrix representation. The figure shows the number of rows of each matrix constructed starting from the patient-related data and the external knowledge. N_{ij} corresponds to the number of nonzero elements in the matrix. Matrix data are available at https://gitlab.com/smarini/MaDDA/tree/master/Patient_similarity_TCGA/matrices.

The resulting matrices (Figure 2) were used to construct the R and Θ block matrices providing the inputs for the factorization algorithm. Note that the R block matrix will be symmetrical except for the R_{36} and R_{63} blocks since these 2 matrices were obtained by using 2 different knowledge sources (ie mutation and gene expression data).

All the object ranks except the patient one were initialized as in Eq. 10, while the patient rank k_6 was initialized to the value resulted by performing a grid search and computing the dispersion coefficient as in Eq. 11.

The convergence of the algorithm was then monitored by measuring the objective function (Eq. 5). The algorithm stopped when the difference between 2 consecutive norms was under the threshold 10^{-5}. The number of repetitions n was set to 10, in order to reduce the effect of the initialization. The resulting consensus matrix $\hat{C}$ was used as a similarity matrix to extract patient-patient similarities.

Validation of real data results

To validate the results on the real data set, we further applied the proposed algorithm on the same data set (Figure 2) but excluding the *Survival* column (ie $n_1 = 3$) that corresponds to what we want to estimate. We applied a hierarchical clustering on the resulting patient similarity matrix where the linkage was determined using complete link method.[55] The survival curves were estimated by the

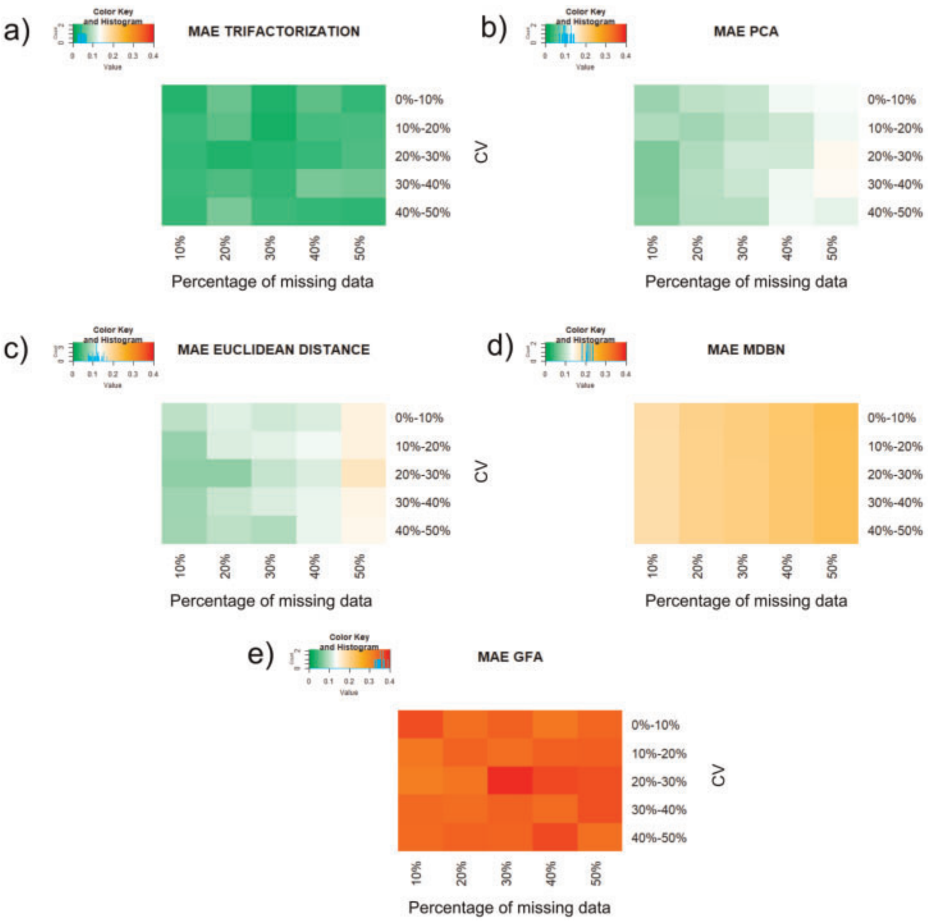


Figure 3. Heatmaps of the mean absolute error (MAE). MAEs were obtained by applying: A, the proposed trifactorization algorithm; B, principal component analysis (PCA); C, Euclidean distance; D, multimodal deep belief network (MDBN); E, group factor analysis (GFA) to the 25 synthetic data sets. The simulated data sets were constructed by considering different percentages of missing data (ie heatmap columns) and noise (ie heatmap row). The noise was added based on the coefficient of variation (CV) (Eq. 12). The map clearly shows smaller MAEs for the proposed approach. The trifactorization gave results less sensitive to sparse data (ie high number of missing data).

Kaplan-Meier method[56] and were compared using the log-rank test.[57]

As for the simulated data sets, we assessed and compared the real data results of the proposed methodology with the ones obtained by applying the PCA, ED, MDBN, and GFA to the patient-related data. As for the trifactorization approach, a hierarchical clustering using complete link method was applied to the similarity matrices resulting by applying these techniques. The different performances were finally evaluated by comparing survival curves of the cluster of patients identified.

RESULTS

Simulation study

To evaluate the performance of the proposed approach, we produced 25 synthetic data sets as described in Simulated data set construction section.

The k parameters (Selection of initial parameters section) were initialized following Eq. 10 for all the objects except the patient rank (k_6). k_6 was defined through a grid search (Eq. 11) on the simulated data set with 10% of missing data and noise between 0% to 10% (Supplementary Material Table S1). $k_6 = 10$ resulted as the rank with the highest values of ρ (Supplementary Material Table S2).

The proposed algorithm was therefore applied to all the 25 data sets by selecting as input $k_6 = 10$, while the other objects ranks were computed according with Eq. 10.

The trifactorization performances on simulated data were evaluated by comparing the algorithm's MAE (Eq. 13) with the MAEs obtained by applying the PCA, ED measure, MDBN, and GFA on the same data sets. The results are shown through heatmaps in Figure 3 (details in Supplementary Material Table S3). These confirmed the best performance are achieved with the trifactorization algorithm, and our approach showed its robustness in presence of different similarity structures and missing data.

Validation study: patient similarity in acute myeloid leukemia

We investigated the patient-patient similarities uncovered by the trifactorization algorithm by applying it to a real case, that is considering the AML data sources reported in Figure 1. The rank parameters were computed as in Eq. 10 for all the objects except the patient 1 that was set to $k_6 = 5$ (ie rank with the highest ρ values—Eq. 11). The other input ranks obtained were: $k_1 = 2$,

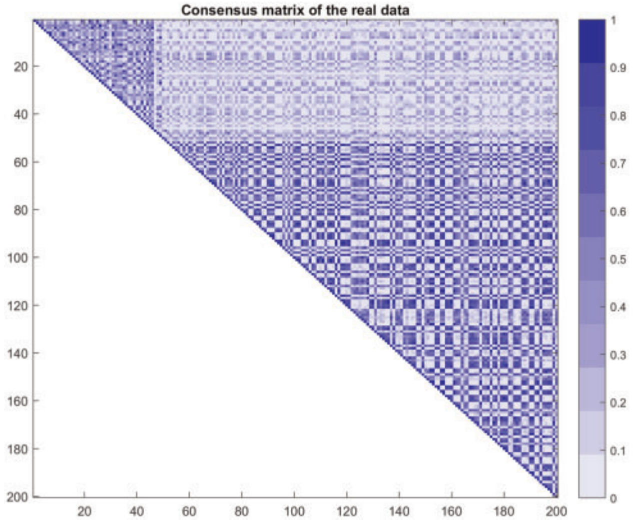


Figure 4. Consensus matrix obtained by applying the trifactorization algorithm to the acute myeloid leukemia (AML) data. Matrix rows and columns correspond to the 200 patients, and it was constructed by considering the resulting G_6 matrix (An overview of the factorization algorithm section). The matrix shows 2 groups of patients clustered together (the corresponding dendrogram is reported in Supplementary Material Figure S1).

$k_2 = 14,\ k_3 = 480,\ k_4 = 13,\ k_5 = 19$. The consensus matrix obtained by applying this procedure is shown in Figure 4.

The result validation was then conducted by applying the proposed algorithm to same data set but not containing the survival information (Result validation section). The trifactorization revealed 2 major groups that we labeled $\mathbf{G_1}$ and $\mathbf{G_2}$ (dendrogram reported in Supplementary Material Figure S1).

In order to validate the obtained clusters, we plotted the Kaplan-Meier survival curves of the 2 patients' groups, as reported in Figure 5A. The survivals of the 2 groups were clearly different (log rank P-value = .0159) with G_1 indicating a better prognosis than G_2.

In addition, we compared our results by performing on the AML data set the PCA, ED, MDBN, GFA methods. In Figure 5 are shown all the survival curves obtained with these approaches and no statistical difference between the 2 groups was found (ED P-value = .7763, PCA P-value = .6278, MDBN P-value = .2954, and GFA P-value = .1652). Taken together, these analyses demonstrated the capacity of the proposed method to add value in integrating different knowledge sources and provide patient-patient similarities for personalized medicine.

DISCUSSION

The availability of increasingly larger amount of biomedical data is pushing researchers and companies to investigate useful information by combining data from difference resources, with the aim of unveiling hidden knowledge on patients, diseases and therapies. However, biomedical data are characterized by high complexity, heterogeneity, sparsity,[58] and "*large m small p*" problems.[59] Sparsity,[58] is a characteristic of matricial/graph representations where most of the elements are null, that is zeros/absent links. This particularly true for genomic data (eg in a gene network, the number of links connecting the nodes will be orders of magnitude smaller than the number of nodes), or temporal data (eg hospitalizations, complications and different blood tests for different patients are sparse along a mostly uneventful time line). "*large m small p*" problems,[59] is the fact that the number of records (predictors) can be orders of magnitude bigger than samples (observations). For example, a gene panel can measure the expression of thousands of genes, but cohorts are usually tens to hundreds of patients.

A number of approaches have been recently proposed to data integration and to bring out intrinsic characteristics of the data. In particular, a specific class of methods rely on the dimension reduction of the data through the definition of metafeatures that allows the projection of the data into a low dimensional space. This property can be used in a machine learning framework in order to capture the hidden interaction effects between variable.

State-of-the-art algorithms are typically based on the metafeature extraction method,[20] or the patient distance measure.[19] They also leverage on a small set of different data types, mostly gene expression,[15–17,19,20,22] methylation,[19,20] copy number variation,[15] protein interactions,[17] clinical data,[16] and diseases.[18] Gligorijevic et al,[21] used also drug information since their aim was to reposition drugs for subgroups of patients.

In this work, we proposed a framework based on matrix trifactorization that integrates several source of data and knowledge with the aim of predicting patient similarity. The proposed approach combines several more data and external knowledge sources respect to other methods recently developed. In detail, our approach provides a comprehensive framework for the prediction of patent similarity by fusing both clinical and multiomics data of patients, and automatically integrating them with external knowledge sources (eg gene-gene interactions, disease-disease associations).

The findings obtained by applying the trifactorization algorithm on synthetic data sets showed better performances if compared to other traditional methods (Figure 3). Moreover, this novel strategy provides more resistance to sparsity and noise, which is ubiquitous in biological data. The application of the trifactorization algorithm to the real AML data set underlined 2 big groups of similar patients (Figure 4). To further confirm this finding, we searched for the exclusive gene signatures characterizing each group, in order to investigate the presence of a molecular mechanism distinguishing the 2 groups. A univariate analysis of gene expression level did not provide significant differences (data not shown). On the other hand, by analyzing mutation data, we extracted a common gene signature consisting of 19 genes mutated in at least one patient in both groups (Supplementary Material Table S4). As expected, these genes resulted significantly enriched for 'leukemia' Online Mendelian Inheritance in Man (OMIM),[60] annotation (Table 3; Fisher's exact P-value = .00245;—enrichment results shown in Supplementary Material Table S5). In addition, we further searched for the genes whose mutation frequency differs significantly between the 2 groups [Fisher's Exact test, False Discovery Rate (FDR) adjusted P-value]. We found 5 genes, namely IDH1 ($P = .00022$), NPM1 ($P = 3.54867\text{e}{-14}$), NRAS ($P = .00145$), PTPN11 ($P = .01$), TET2 ($P = .00015$). Mutations associated with all these genes have been found to be highly related to AML development and progression.[61–65] These findings confirm both the 2 identified groups characterized by a common AML gene signature. Finally, we identified gene signatures discriminating the 2 groups by retrieving the mutations present exclusively either in patients of G_1 (better prognosis) or the G_2 (good prognosis). The gene signatures resulted in 342 (Supplementary Material Table S6) and 52 (Supplementary Material Table S7) genes for G_1 and G_2 group, respectively. An OMIM enrichment analysis conducted on the 2 gene signatures retrieved the term 'leukemia' only

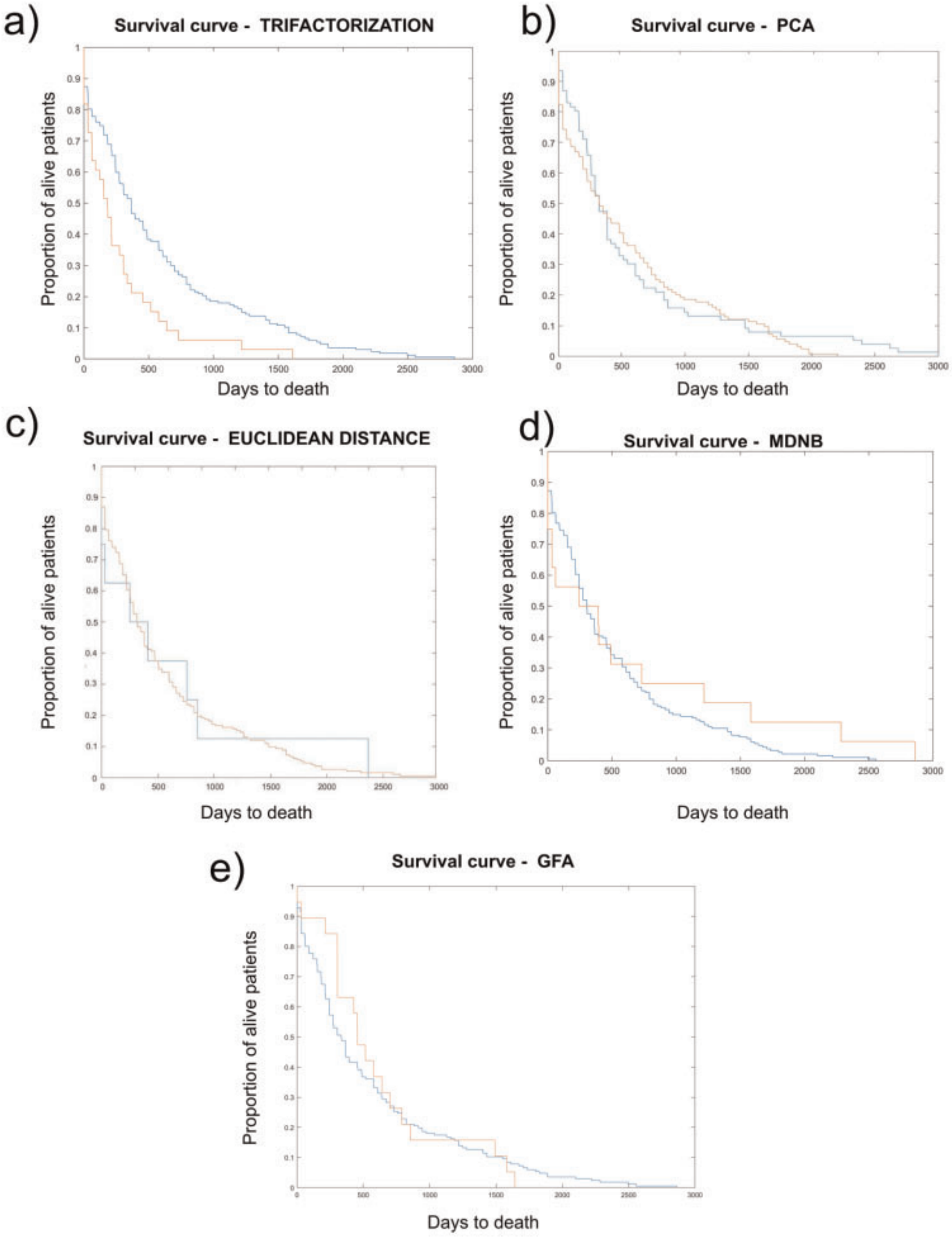


Figure 5. Survival curves. Survival curves corresponding to the clusters obtained by the trifactorization algorithm (A), principal component analysis (PCA) (B), Euclidean distance (ED) measure (C), multimodal deep belief network (MDBN) (D), and group factor analysis (GFA) (E). All the plots report the *P*-value (*P*) resulting from the log-rank test. Only the 2 clusters obtained with the proposed approach have statistically significant survival curves (trifactorization *P*-value = .01; ED *P*-value = .7763; PCA *P*-value = .6278; MDBN *P*-value = .2954; GFA *P*-value = .1652).

for G_2 (Table 3; Fisher's exact *P*-value = .0175, complete results in Supplementary Material Tables S8 and S9), coherently with the hypothesis the second group, associated with a poor survival prognosis, presents a more aggressive instance of AML. Interestingly, the groups are not significantly different for sex (*P* = .2594, Fisher's Exact test) or age (*P* = .0884, Wilcoxon rank sum test), indicating that our unsupervised method did not simply discriminated obvious patient subgroups, but was able to mine more subtle differences characterizing poor-vs-good prognoses.

The current study showed is power in discovery patient-patient similarity by integrating several data and knowledge sources that involve associations between clinical data, mutations, genes, diseases, pathways, and patients to extract patient similarity. The same approach can be applied to classify novel patients and, the results might be used to suggest potential tailored treatments based on successful drug treatments extracted from the patient's clinical histories. Moreover, further improvements and novel features can be provided by directly including into the trifactorization framework data sources involving drugs (ie DrugBank,[66] PharmGKB,[67]). In this way, by selecting as target the patient-drug matrix, the algorithm could predict targeted therapies for specific profiles of patients.

Table 3. OMIM enrichment results

OMIM term	Significant *P*-value (Fisher's exact test)	OMIM enrichment analysis of the genes associated with
Leukemia	.00246	Common mutations between G1 and G2
Bardet-biedl_syndrome	.026431	Mutations of G1
Macular_degeneration	.037245	
Leukemia	.017554	Mutations of G2
Long_qt_syndrome	.030766	
Fibrosis	.030766	
Cone-rod_dystrophy	.038311	
Thyroid_carcinoma	.043309	

Note: Significant *P*-values ($P < .05$). The full enrichment results for the genes associated with common mutations between G1 and G2, mutations of G1, and mutations of G2 are respectively shown in Supplementary Material Tables S5, S8, and S9.

OMIM: Online Mendelian Inheritance in Man.

The main limitations of the proposed approach are 2-fold:

- On the one hand, it is necessary to represent data and knowledge in terms of bidimensional relation matrices. This often requires flattening concept hierarchies and graphs, thus generating matrices of very high dimensionality. The computational burden of the optimization algorithm is highly dependent on the number of data sources and on the size of the relational matrices, in particular when such matrices are not sparse.
- On the other hand, the method has a number of crucial design parameters, the ranks, which requires fine tuning in order to obtain a compromise between the quality of matrix reconstruction and the need of representing data with low dimensionality latent vectors.

Nevertheless, we believe that the proposed approach represents a powerful and interesting strategy to deal with data and knowledge fusion for similarity calculation, which may provide advantages in precision oncology applications.

Finally, the proposed approach is particularly suitable for precision oncology due to the high number of available data sources on cancer, but it can be easy applied to other fields and diseases.

CONCLUSION

In this work, we analyzed the problems related to the application of precision oncology. Literature evidence seems to show how a traditional clinical trial approach looking for biomarker-specific patient subgroups is particularly hard to implement due to heavy, intrinsic statistical limitations. We have shown how the problem of subgroup identification can be solved with a data fusion approach exploiting relations of heterogeneous, multidimensional data sources. In our application, we considered objects as diverse as clinical data, diseases, genes, mutations, pathways, and the patients themselves. Thanks to a latent feature representation via matrix trifactorization, we were able to identify clinically meaningful patient subgroups. The approach showed better performance when compared with standard clustering strategies. Future works will deal with optimizing the factorization strategy and to provide automated explanations of the results obtained.

FUNDING

This work was supported by the 3 project grant no. 2015-0042 "Genomic profiling of rare hematologic malignancies, development of personalized medicine strategies, and their implementation into the Rete Ematologica Lombarda (REL) clinical network."

Conflict of interest statement. None declared.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We thank Ivan Limongelli for the help in organizing genome data. We are also grateful to Marco Piastra, Gianluca Gerard and Stefano Montoli for the analysis carried on with MDBN. We sincerely acknowledge Blaz Zupan and Marinka Zitnik for their original ideas, and for their support and training in applying trifactorization methods.

REFERENCES

1. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015; 372 (9): 793–5.
2. Lu YF, Goldstein DB, Angrist M, Cavalleri G. Personalized medicine and human genetic diversity. *Cold Spring Harbor Perspect Med* 2014; 4 (9): a008581.
3. Chin L, Gray JW. Translating insights from the cancer genome into clinical practice. *Nature* 2008; 452 (7187): 553–63.
4. Sparano JA, Paik S. Development of the 21-gene assay and its application in clinical practice and clinical trials. *J Clin Oncol* 2008; 26 (5): 721–8.
5. Parker JS, Mullins M, Cheang MC. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* 2009; 27 (8): 1160–7.
6. Pellagatti A, Benner A, Mills KI, *et al*. Identification of gene expression-based prognostic markers in the hematopoietic stem cells of patients with myelodysplastic syndromes. *J Clin Oncol* 2013; 31 (28): 3557–64.
7. Meric-Bernstam F, Brusco L, Shaw K, *et al*. Feasibility of large-scale genomic testing to facilitate enrollment onto genomically matched clinical trials. *J Clin Oncol* 2015; 33 (25): 2753–62.
8. Group E-ACR. Executive Summary: Interim Analysis of the NCI-MATCH Trial. Secondary Executive Summary: Interim Analysis of the NCI-MATCH Trial. 2016. http://ecog-acrin.org/nci-match-eay131/interim-analysis. Accessed June 2016.
9. Tredan O, Corset V, Wang Q, *et al*. Routine molecular screening of advanced refractory cancer patients: An analysis of the first 2490 patients of the ProfiLER study. *J Clinical Oncol* 2017; 35 (18_suppl): LBA100.
10. Le Tourneau C, Delord JP, Goncalves A, *et al*. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. *Lancet Oncol* 2015; 16 (13): 1324–34.

11. Prasad V, Vandross A. Characteristics of exceptional or super responders to cancer drugs. *Mayo Clin Proc* 2015; 90 (12): 1639–49.
12. Biankin AV, Piantadosi S, Hollingsworth SJ. Patient-centric trials for therapeutic development in precision oncology. *Nature* 2015; 526 (7573): 361–70.
13. Sun J, Wang F, Hu J, Edabollahi S. Supervised patient similarity measure of heterogeneous patient records. *ACM SIGKDD Explor Newsl* 2012; 14 (1): 16–24.
14. Brown SA. Patient similarity: emerging concepts in systems and precision medicine. *Front Physiol* 2016; 7: 561.
15. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics (Oxford, England)* 2009; 25 (22): 2906–12.
16. Ow GS, Tang Z, Kuznetsov VA. Big data and computational biology strategy for personalized prognosis. *Oncotarget* 2016; 7 (26): 40200–20.
17. Xu T, Le TD, Liu L, Wang R, Sun B, Li J. Identifying cancer subtypes from miRNA-TF-mRNA regulatory networks and expression data. *PLoS One* 2016; 11 (4): e0152792.
18. Girardi D, Wartner S, Halmerbauer G, Ehrenmüller M, Kosorus H, Dreiseitl S. Using concept hierarchies to improve calculation of patient similarity. *J Biomed Inform* 2016; 63: 66–73.
19. Wang B, Mezlini AM, Demir F, *et al*. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* 2014; 11 (3): 333–7.
20. Liang M, Li Z, Chen T, Zeng J. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM Trans Comput Biol and Bioinf* 2015; 12 (4): 928–37.
21. Gligorijevic V, Malod-Dognin N, Przulj N. Patient-specific data fusion for cancer stratification and personalised treatment. *Pac Symp Biocomput* 2016; 21: 321–32.
22. Planey CR, Gevaert O. CoINcIDE: a framework for discovery of patient subtypes across multiple datasets. *Genome Med* 2016; 8 (1): 27.
23. Wang F, Tao L, Changshui Z. Semi-supervised clustering via matrix factorization. In: *Proceedings of the 2008 SIAM International Conference on Data Mining*. Atlanta, GA: SIAM; 2008.
24. Zhang P, Wang F, Hu J. Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity. In: *AMIA Annual Symposium Proceedings*. Bethesda, MD: American Medical Informatics Association; 2014.
25. Zitnik M, Janjic V, Larminie C, Zupan B, Przulj N. Discovering disease-disease associations by fusing systems-level molecular data. *Sci Rep* 2013; 3 (1): 3202.
26. Zitnik M, Nam EA, Dinh C, Kuspa A, Shaulsky G, Zupan B. Gene prioritization by compressive data fusion and chaining. *PLoS Comput Biol* 2015; 11 (10): e1004552.
27. Zitnik M, Zupan B. Matrix factorization-based data fusion for gene function prediction in baker's yeast and slime mold. *Pac Symp Biocomput* 2014; 19: 400.
28. Žitnik M, Zupan B. Matrix factorization-based data fusion for drug-induced liver injury prediction. *Syst Biomed* 2014; 2 (1): 16–22.
29. Vitali F, Cohen LD, Demartini A, *et al*. A network-based data integration approach to support drug repurposing and multi-target therapies in triple negative breast cancer. *PLoS One* 2016; 11 (9): e0162407.
30. Zitnik M, Zupan B. Data fusion by matrix factorization. *IEEE Trans Pattern Anal Mach Intell* 2015; 37 (1): 41–53.
31. Singh AP, Gordon JG. Relational learning via collective matrix factorization. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Las Vegas, NV: ACM; 2008.
32. Klami A, Virtanen S, Leppäaho E, Kaski S. Group Factor Analysis. *IEEE Transactions on Neural Networks and Learning Systems* 2015; 26 (9): 2136–47.
33. Ruffini M, Gavalda R, Limon E. Clustering patients with tensor decomposition. In: Finale D-V, Jim F, David K, Rajesh R, Byron W, Jenna W, eds. *Proceedings of the 2nd Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research: PMLR*. Boston, MA: PMLR; 2017: 126–46.
34. Khan SA, Leppäaho E, Kaski S. Bayesian multi-tensor factorization. *Mach Learn* 2016; 105 (2): 233–53.
35. Virtanen S, Klami A, Khan AK, Kaski S. Bayesian group factor analysis. In: Artificial Intelligence and Statistics. La Palma, Canary Islands: AISTATS; 2012.
36. Klami A, Virtanen S, Leppäaho E, Kaski S. Group factor analysis. *IEEE Trans Neural Netw Learn Syst* 2015; 26 (9): 2136–47.
37. Wang Y-X, Zhang Y-J. Nonnegative matrix factorization: a comprehensive review. *IEEE Trans Knowl Data Eng* 2013; 25 (6): 1336–53.
38. Pinero J, Bravo A, Queralt-Rosinach N, *et al*. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* 2017; 45 (D1): D833–d39.
39. Hudson TJ, Anderson W, Aretz A, *et al*. International network of cancer genome projects. *Nature* 2010; 464 (7291): 993.
40. Chatr-Aryamontri A, Breitkreutz B-J, Oughtred R, *et al*. The BioGRID interaction database: 2015 update. *Nucleic Acids Res* 2014; 43 (D1): D470–D78.
41. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000; 28 (1): 27–30.
42. Kibbe WA, Arze C, Felix V, *et al*. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res* 2014; 43 (D1): D1071–D78.
43. Gao J, Aksoy BA, Dogrusoz U, *et al*. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013; 6 (269): pl1.
44. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 2003; 31 (4): e15.
45. Limongelli I, Marini S, Bellazzi R. PaPI: pseudo amino acid composition to score human protein-coding variants. *BMC Bioinformatics* 2015; 16 (1): 123.
46. Rappaport N, Nativ N, Stelzer G, *et al*. MalaCards: an integrated compendium for diseases and their annotation. *Database* 2013; 2013: bat018.
47. Cokelaer T, Pultz D, Harder LM, Serra-Musach J, Saez-Rodriguez J. BioServices: a common Python package to access biological Web Services programmatically. *Bioinformatics* 2013; 29 (24): 3241–2.
48. Brown CE. *Coefficient of Variation. Applied Multivariate Statistics in Geohydrology and Related Sciences*. Berlin, Heidelberg: Springer Berlin Heidelberg; 1998: 155–57.
49. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geosci Model Dev* 2014; 7 (3): 1247–50.
50. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemom Intell Lab Syst* 1987; 2 (1-3): 37–52.
51. Hinton G. A practical guide to training restricted Boltzmann machines. *Momentum* 2010; 9 (1): 926.
52. Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. *Neural Comput* 2006; 18 (7): 1527–54.
53. Lowenberg B, Downing JR, Burnett A. Acute myeloid leukemia. *N Engl J Med* 1999; 341 (14): 1051–62.
54. Dohner H, Weisdorf DJ, Bloomfield CD. Acute myeloid leukemia. *N Engl J Med* 2015; 373 (12): 1136–52.
55. Hartigan JA, Hartigan J. *Clustering Algorithms*. Wiley: New York; 1975.
56. Dinse GE, Lagakos SW. Nonparametric estimation of lifetime and disease onset distributions from incomplete observations. *Biometrics* 1982; 38 (4): 921–32.
57. Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Ann Stat* 1988; 16: 1141–54.
58. Ye J, Liu J. Sparse methods for biomedical data. *SIGKDD Explor Newsl* 2012; 14 (1): 4–15.
59. Scott DW. The curse of dimensionality and dimension reduction. In: *Multivariate Density Estimation: Theory, Practice, and Visualization*. 2nd ed. New York: Wiley; 1992: 217–40.
60. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* 2015; 43 (D1): D789–98.

61. Paschka P, Schlenk RF, Gaidzik VI, *et al*. IDH1 and IDH2 mutations are frequent genetic alterations in acute myeloid leukemia and confer adverse prognosis in cytogenetically normal acute myeloid leukemia with NPM1 mutation without FLT3 internal tandem duplication. *J Clin Oncol* 2010; 28 (22): 3636–43.
62. Verhaak RG, Goudswaard CS, van Putten W, *et al*. Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. *Blood* 2005; 106 (12): 3747–54.
63. Schlenk RF, Dohner K, Krauter J, *et al*. Mutations and treatment outcome in cytogenetically normal acute myeloid leukemia. *N Engl J Med* 2008; 358 (18): 1909–18.
64. Bentires-Alj M, Paez JG, David FS, *et al*. Activating mutations of the noonan syndrome-associated SHP2/PTPN11 gene in human solid tumors and adult acute myelogenous leukemia. *Cancer Res* 2004; 64 (24): 8816–20.
65. Gaidzik VI, Paschka P, Spath D, *et al*. TET2 mutations in acute myeloid leukemia (AML): results from a comprehensive genetic and clinical analysis of the AML study group. *J Clin Oncol* 2012; 30 (12): 1350–7.
66. Law V, Knox C, Djoumbou Y, *et al*. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* 2014; 42 (D1): D1091–7.
67. Hewett M, Oliver DE, Rubin DL, *et al*. PharmGKB: the pharmacogenetics knowledge base. *Nucleic Acids Res* 2002; 30 (1): 163–65.